Universitat de Lleida

Prediction of the User's political trends with Twitter: Silvereye Tool

MARC SOLÉ FARRÉ

Directors: FRANCESC GINÉ DE SOLA and MAGDA VALLS MARSAL

Polytechnic School - Universitat de Lleida

18-04-2017

Contents

Acknowledgement

” M’agradaria aprofitar aquest petit racó del document per deixar constància i agrair a totes les persones que m’han donat suport i ajudat durant el transcurs de la carrera, màster i projecte. Per començar m’agradaria agrair l’ajuda i suport a tota la meva família, des dels meus pares fins als meus tiets, padrins i cosines. Tots ells han estat presents en aquest viatge i mai han dubtat amb ajudar-me sempre que ho he necessitat. Els meus pares, Antonio Solé i Antonieta Farré, que són els que sempre m’han fet costat amb totes les decisions que he pres i que tota la vida han aguantat les meves penes i alegries, gràcies per la vostra paciència i dedicació. També amb els meus tiets, Josep Farré i Josep Mª Solé i Concepció Pont, i cosines Yolanda Solé i Vanessa Solé per confiar amb mi, i animar-me. Als meus padrins Josep Solé i Mª Tresa Claria que tot i que les noves tecnologies no son el seu punt fort i no entenen moltes de les conversacions que tenia amb ells sobre el projecte, sempre m’han escoltat i m’han aconsellat amb la seva frase de, ”tu treballa i no et preocupis”. Recordar als meus padrins difunts, Josefina Coll i Josep Farré que tot i que ja no hi són sempre han estat un referent per mi i mai oblidaré. També vull donar les gràcies per la paciència que han tingut els companys de pis Gabriel Lozano i Marc Saladrigues, que m’han acompanyat duran tota la carrera i màster, i que hem intercanviat coneixements i idees.

En especial m’agradaria donar els meus agraïments als meus directors del projecte Frances Giné i Magda Valls, per la seva inestimable ajuda i dedicació, ja que han dedicat molta paciència i hores amb aquest projecte i a mi.

Gràcies al grup d’investigació Distributed Computing Group (GCD) i the Cryptography & Graphs Research Group (C&G) per brindar-me i compartir coneixements.

Donar les gràcies també als companys de despatx, Marc Gonzalez, Eloi Gabaldon i Sergi Vila pels consells, ajudes i converses que hem tingut.

Em faltarien dies i pàgines per donar els agraïments a totes les persones que han estat presents aquests anys i han col·laborat directament i indirectament al projecte i durant el transcurs de la carrera, amics, companys de classe, professors, companys del GDG, i més. Moltes gràcies a tots i totes per aportar-me el vostre granet de sorra al llarg d’aquests anys.

Gràcies. ”

List of Figures

List of Tables

Chapter 1

Context of the project

In this chapter, we will describe the framework of the project together with its motivation. Furthermore, the main objectives will be explained.

1.1 Introduction

With augmentation in the use of the social networks and the increasing quantity or amount of content uploaded and published, it can be found a great deal of information about the users and their tendency. This occurs because most of the content of these socials networks is formed by the opinions and the interest of its users [1]. So, nowadays, the social networks have turned into an excellent platform to analyze the social trends in different fields, such as sport, fashion, politics [2].

Specifically, we are interested in forecasting the political trends of Twitter users. We can found several examples where the traditional forecast methods, such as telephone polls which do not work; which gives an idea about the difficulty to predict the results of the political elections in our society. As a representative example, we have the case of the Spanish elections in 2016 [3], when the results of the polls showed a significant difference from the real results. In some cases the differences between the real results and the polls results were 20 points. So, the prediction of the results of a political election using the opinion of the people through the social networks is a challenge for the researching community [4]. Among the three most used social networks: Facebook, Twitter and Instragram. Therefore, we are interested to start experimenting with Twitter because this social network is based on expanding the content shared and most of the content uploaded by their users are their opinions. On Facebook, we can also find users opinions and the sharing of content is more restricted for the user friends. Also, Instagram is a social network based on expanding the content shared but the content is more related to lifestyle and it does not contain many opinions of its users.

In this project, we want to group the users according to collections. Which, in our case are the political parties. In other words, this project is focused on defining two strategies to analyze the users of political tendency through Twitter. The first one is based on analyzing the users tweets by means of discovering the sentiment of the tweets and classifying them according to the political party that they support. The other strategy is based on analyzing and exploiting the relations between users. Accordingly, these relationships can be represented by means of a graph structure. This structure can be used to cluster the users in appropriate groups related to political parties.

This report explains the design, development and implementation of both strategies, including used technologies, architecture and the problems found in the development.

1.2 Objectives

The main aim of this project is to develop the SilverEye tool to analyze Twitter users in order to forecast their political trends. According to this, the SilverEye tool will integrate both strategies explained above.

The objectives to achieve are:

- Getting knowledge and applying big data tools to manage tweets in real time.
- Designing an infrastructure to extract and manipulate twitter data in real time.
- Analyzing and classify according to the collections they belong.
- Designing and implementing both strategies to forecast political trends.
- Showing the trend of a set of users according to some parameters such as location, sex or age.
- Visualizing the results in a web platform.
- Comparing the efficiency of both strategies in terms of accuracy of the prediction and processing time.

To achieve these goals, we have to tackle the following problems:

- The selection and subsequent knowledge of the suitable Big Data Tools.
- The number of political tweets by user may be poor to infer correctly its tendency.
- The correct wording of the tweets may difficult its correct analysis.
- Very short tweet messages give us poor information to be analyzed.
- The irony in tweets can also difficult their analysis.
- In the political context we can have some daily political facts that influence to sentiment analysis.
- The use of external links in a tweet message to complement what the user wants to express has to be handled.

1.3 Structure

This report is organized in six chapters. Chapter 1 is dedicated to providing a brief introduction of the project and its main goals. In Chapter 2, we overview the state of the art, on forecasting political elections results using the social networks and we discuss some of the present solutions. Chapter 3 presents the related technologies used for developing the proposed strategies. Chapter 4 presents the strategies and its development. Chapter 5 shows the results of the strategies and finally the Chapter 6 describes the main conclusions of the project. Finally, there is an appendix which provides some illustrative examples to complete the overall presentation.

Chapter 2

State of the art

In this chapter of art, we introduce the solutions provided by the literature to forecast the results of the political elections by means of analyzing the social network data.

2.1 Present solutions

Nowadays, the most popular system of forecasting the results of the political elections is based on doing survey forms [5], but sometime they are not very accurate [3] and also have an expensive cost and long time to make the results. In the last years, the social networks have awaken the interest of the researcher community [6, 7] to predict the result of the political elections because it gives every day information about the social life of the users and it can show the change of the tendencies in real time.

2.2 Forecasting political election results

In literature, we can find several extensions of the forecasting political election results like the works of [8–10]. Special interest has the work of Tumasjan et all in [4], who analyzed 100,000 Twitter messages mentioning parties or politicians prior to the German federal election of 2009. Such as they present, the mere number of tweets reflects voter references and comes close to traditional election polls, while the sentiment of Twitter messages closely corresponds to political programs, candidate profiles, and evidence from the media coverage of the campaign trail. According to the results presented in this paper, they demonstrate that Twitter can be seen as a valid real-time indicator of political sentiment. Likewise in the future work, they talked about the unstructured nature of microblogging communication and therefore include searchable keywords, so-called hashtags, in many messages (e.g., #PP, #PSOE). Thus, these hashtags include these missing pieces of information. Another thing that they talked about is the possible information that twitter relationships can give. In fact, this work inspires us to make the strategies presented in our work. In addition, in contrast of them, we want to analyze the users profiles content, in particular, the tweets and relationships.

2.3 Sentiment Analysis

There are numerous research papers and studies that focus on sentiment classification for Twitter, like [11–13]. These studies describe some interesting methodologies of detecting and identifying sentiment from twitter data. We found different techniques to identify the Sentiment analysis such as the Machine Learning Approach (ML) which uses linguistic features. The Lexicon-based Approach that relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which uses statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. We have used some of the methodology described in [13].

Which uses emoticons for Tweets feature extraction and including the usage of unigrams, bigrams + unigrams and parts of speech tags for making a sentiment algorithm to classify the sentiment of the tweets. This paper inspired me to make the sentiment analysis algorithm.

2.4 Tweet classification

The main goal of tweets classification is to integrate all tweets that are related semantically in one place on the Web. The data inequality problems exists because authors post tweets and hashtag based on their own style, without any kind of pattern. For instance, the authors can post tweets with the same hashtags in different languages (English - Arabic). Therefore, the hashtags cannot be the only trackback function of the tweet to link it to the related tweets. Ghaly et all in [14] presents an automatic tweet classifier, which is not limited by the language or the category of the tweet. The approach can create new suggested multilingual (English, Arabic) hashtags from tweet's content and comments "auto- tagging" in social linked open data (LOD), matching the tweets in the same topic. Thus, it provides high accuracy and performance using blocking and indexing techniques. This has inspired us to improve the tweet classifier and for showing information about the main and related political themes used in twitter in relation to political parties.

Chapter 3

Related technologies

This chapter is dedicated to present the technologies used to develop this project. It is organized according to the scope action of the different tools. Hence in the Big Data Tools section, we find the tools used for managing and processing efficiently large data sets. In the Twitter API section we can find the tools used to extract twitter data, as well as to store data, such as the case of Mongodb. Finally, the last section covers the tool for data visualization and management.

3.1 Big data tools

In the next sections, will we describe the tools used for extracting and transforming data. In the project, we combine several tools.

3.1.1 Hadoop

The Apache Hadoop [15] is an open-source software framework designed to allow distributed processing of large data sets across clusters of computers.

A typical Hadoop cluster includes one master node and multiple slave nodes. While the aim of the master is job tracking, task tracker, name node management and data node management, the aim of slave nodes is the storage of data, the task on computation and the corresponding task tracker. In Figure 3.1, we can see the node architecture of the Hadoop cluster.

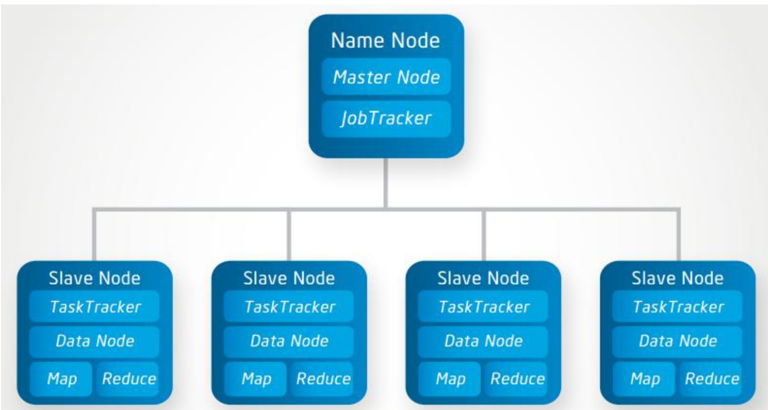

Figure 3.1: Hadoop node architecture

Hadoop allows the MapReduce [16] programming model which consists on splitting or mapping the data to a blocks, executing simple operations for each blocks and finally joining this with the reduce operation. In addition Hadoop cluster provides a MapReduce engine with a job tracker in order to the client applications send MapReduce jobs.

One interesting attribute of Hadoop is the Distributed File System (HDFS) [17] , which is scalable and portable, where every node has an unique data block. Therefore, Hadoop has a HDFS data cluster. Every node serves data blocks over the network using a block protocol specific of HDFS. In Figure 3.2 we can see an example of Big Data input which has been split to the HDFS system for the map operation and joined for the reduce operations.

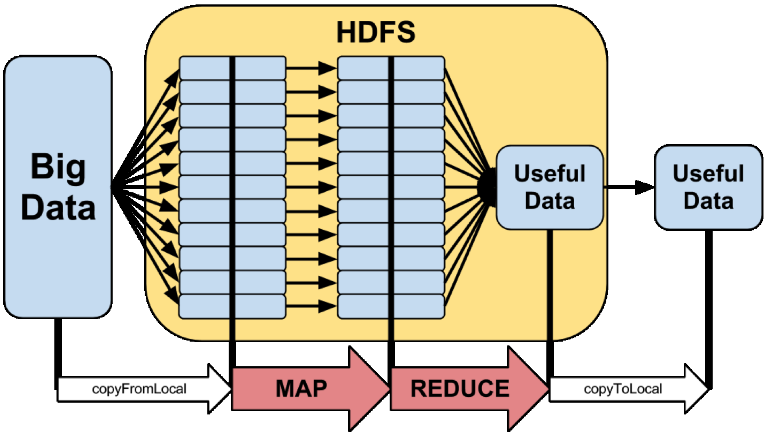

Figure 3.2: MapReduce Work flow

We have selected Hadoop for the MapReduce engine, because it provides us a great performance in the parallelization of the operations and the optimization of the resources. As we will discuss in the development section 4.3.5, the aim of the usage is to take the repeated words and count its appearance in all of the tweets text.

3.1.2 Spark

Apache Spark [18] is also an open-source software framework for massive data processing. This is a fast and general engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. In relation to Hadoop, it extends its MapReduce model to obtain a faster execution and more analysis scenarios; for example, interactive queries and streaming processing. Spark is more efficient than Hadoop because Spark handles most of its operations “in memory” – copying them from the distributed physical storage into far faster logical RAM memory

Apache Spark provides an application programming interface centered on a data structure, named Resilient Distributed Dataset (RDD), a read-only multiset of data items distributed over a cluster machines, which is maintained in a fault-tolerant way.

The availability of RDDs [19] facilitates the implementation of both iterative algorithms followed by the MapReduce model, that visit their dataset multiple times in a loop, and interactive/exploratory data analysis. Thus latency of such applications (compared to Apache Hadoop, a popular MapReduce implementation) may be reduced by several orders of magnitude [18].

Likewise Apache Spark requires a cluster manager and a distributed storage system. For cluster management, Spark supports standalone (native Spark cluster), Hadoop YARN, or Apache Mesos.

Figure 3.3 shows the architecture of Spark. Next, we will describe each of the main parts.

Apache Spark

The most important part of the Spark architecture is Spark Core. This provides distributed task dispatching, scheduling and basic I/O functionalities, centered on the RDD abstraction

Spark SQL

Spark SQL is a component on top of Spark Core that introduced a data abstraction called DataSets, which provides support for structured and semi-structured data. Spark SQL provides a Domain-Specific Language (DSL).

Spark streaming

Spark Streaming leverages Spark Core's on fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD transformations on those mini-batches of data. This design enables the same set of application code written for batch analytics to be used in streaming analytics, thus facilitating easy implementation of Lambda architecture [20].

MLlib machine learning library

Spark MLlib [21] is a distributed machine learning framework on top of Spark Core. Many common machine learning and statistical algorithms have been implemented and are shipped with MLlib which simplifies large scale machine learning pipelines, including:

- Summary statistics, correlations, stratified sampling, hypothesis testing, random data generation.
- Classification and regression: support vector machines, logistic regression, linear regression, decision trees, naive Bayes classification
- Collaborative filtering techniques including Alternating Least Squares (ALS)
- Cluster analysis methods including k-means, and Latent Dirichlet Allocation (LDA)
- Dimensionality reduction techniques such as Singular Value Decomposition (SVD), and Principal Component Analysis (PCA)
- Feature extraction and transformation functions
- Optimization algorithms such as stochastic gradient descent, limited-memory BFGS (L-BFGS)

These wide range of algorithm toolkits, along with the tools for data frame managing, make Spark a interesting tool for this Big Data project.

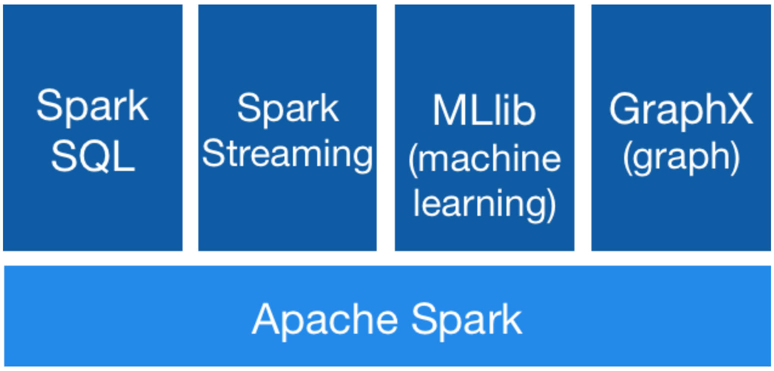

Figure 3.3: Spark Stack

GraphX

GraphX is a distributed graph processing framework on top of Apache Spark. Because it is based on RDDs, which are immutable, graphs are also immutable and thus GraphX is unsuitable for graphs that need to be updated, let alone in a transactional manner like a graph database.

GraphFrame

GraphFrames [22] supports general graph processing, similar to Apache Spark's GraphX library. However, GraphFrames are built on top of Spark DataFrames. It provides high-level APIs in Scala, Java, and Python. It aims to provide both the functionality of GraphX and extended functionality taking advantage of Spark DataFrames. This extended functionality includes Motif finding, DataFrame-based serialization, and highly expressive graph queries.

One of the strategies that we present in this project is based on representing data in a graph and making operations in this, GraphFrames provides implemented algorithms based on graphs which are optimized for begin executed in a cluster. We choose GraphFrame because it's newer, faster, more powerful, and easier to use [23]. Figure 3.4 shows the general architecture of GraphFrame.

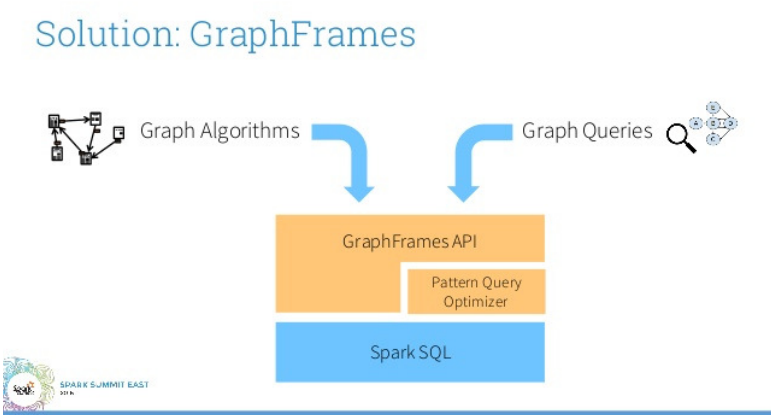

Figure 3.4: GraphFrame Architecture

3.1.3 Storm

Apache Storm [24] is a free and open source distributed realtime computation system. Storm makes it easy to reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. Storm is simpler than Hadoop, since it can be used with any programming language.

Storm has many use cases: Realtime analytics, Online machine learning, Continuous computation, Distributed RPC, ETL, and more. Storm is fast, scalable, fault-tolerant, and it is easy to set up and operate.

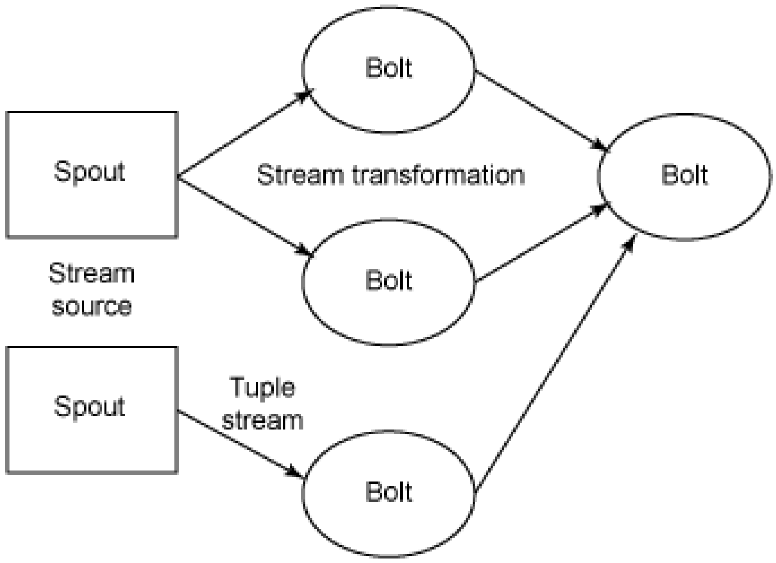

Figure 3.5: Storm Topology

In Figure 3.5, we can see an example of Storm topology, composed by entry data element named Spout, while the Bolts are the elements that transform the data.

A Spout is a source of streams in a topology. Generally Spouts will read tuples from an external source and emit them into the topology (e.g. a Kestrel queue or the Twitter API). Spouts can either be reliable or unreliable. A reliable Spout is capable of replaying a tuple if it failed to be processed by Storm, whereas an unreliable Spout forgets about the tuple as soon as it is emitted.

All processing in topologies is done in Bolts. Bolts can do anything from filtering, functions, aggregations, joins, talking to databases, and more.

Bolts can do simple stream transformations. Doing complex stream transformations often requires multiple steps and thus multiple bolts. For example, transforming a stream of tweets into a stream of trending images requires at least two steps: a Bolt to do a rolling to count of retweets for each image, and one or more Bolts to stream out the top X images (you can do this particular stream transformation in a more scalable way with three bolts than with two).

In order to define our architecture, we have selected Storm to get the data in streaming and manage this data in runtime. It provides an architecture based in modules for extraction and transformation, in addition this is great to disconnect the different operations and test new transformations or add new transformation without manipulating the others. In relation to the data extraction, it provides the possibility of connecting with different data origins.

3.2 Twitter API

Twitter API is the official API to extract information from Twitter application. The libraries used to call the Twitter API is Tweepy [25] for Python and twitter4j [26] for Java. Twitter API gives us different ways to get its data. For our purpose we use Twitter REST API for getting specific information of the users and Streaming Twitter API to get streaming flux of tweets. In the next subsections, we will describe some matters about this.

3.2.1 Twitter REST API

The REST API [27] provides programmatic access to read and write Twitter data. It allows to create a new Tweet, read user profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth [28] and the responses are in JSON format [29].

The amount of requests that Twitter REST API admits is limited. The table of the limitations can be seen in [30]. For instance, if a method allows 15 requests per rate limit window, then it allows you to make 15 requests per window — on behalf of your application. Rate limits are divided into 15 minute intervals. All endpoints require authentication, so there is no concept of unauthenticated calls and rate limits.

In the next Figure 3.6, you can see an example of the different states of one call to the REST API. For example, if we consider a web application which accepts user requests to Twitter REST API, makes one or more requests to Twitter's API. Next it formats and prints the result to the user, as a response to the user's initial request.

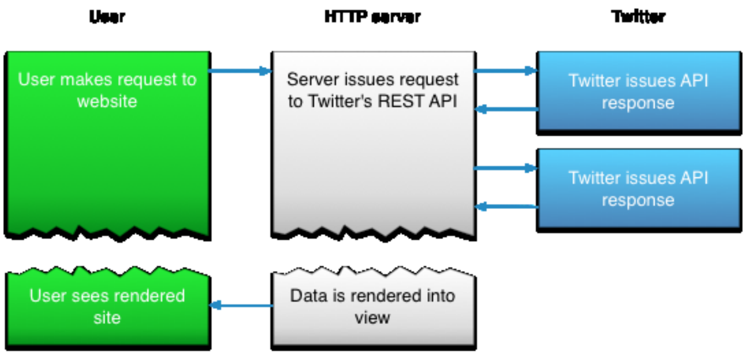

Figure 3.6: Twitter REST APIs Flow

3.2.2 Twitter streaming APIs

The Streaming APIs [31] give low latency access to Twitter's global stream of Tweet data. A streaming client will be able to get Tweets of its interest, according to a set of keywords.

Connecting to the streaming API requires to keep a persistent HTTP connection open. In many cases, this involves thinking about your application differently than if you were interacting with the REST API.

An app which connects to the Streaming APIs will not be able to establish a connection in response to a user request, as shown in the above example 3.6, the client wait for a infinite data flux. Instead, the code for maintaining the Streaming connection is typically run in a process separated from the main process, which handles HTTP requests, as we can see in Figure 3.7.

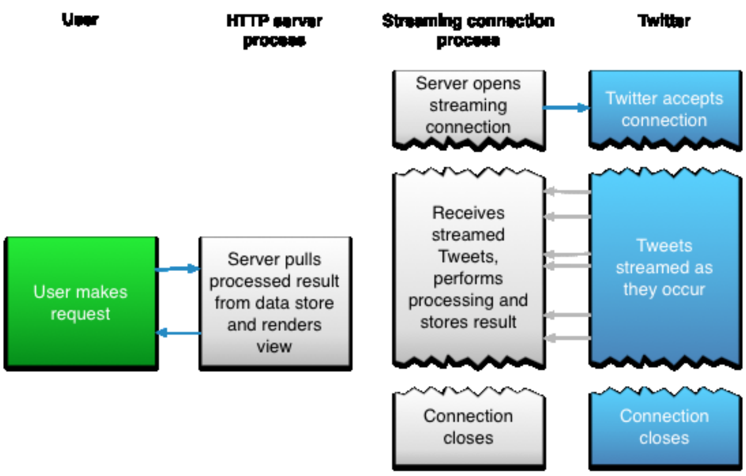

Figure 3.7: Twitter Streaming APIs Flow

3.3 Data storage

3.3.1 Mongodb

MongoDB [32] is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas.

Any relational database has a typical schema design that shows number of tables and the relationship between these tables. In contrast to them, there is no concept of relationship in MongoDB.

We select this database for its flexibility, which supposes a significant advantage in a Bigdata project [33]

3.4 Sentiment Analysis

In this section we present OpeNER, thats is the first tool that we use for analyze and obtain the sentiment of a tweet. The sentiment can be positive, negative or neutral.

3.4.1 OpeNER

OpeNER is a language analysis toolchain to help (academic) researchers and companies make sense out of "natural language analysis" it provides the following utilities:

1. Detect the language of a text.
2. Tokenize texts.
3. Determine polarisation of texts (sentiment analysis) and detect what topics are included in the text.
4. Detect entities named in the texts and link them together. (e.g. President Obama or The Hilton Hotel)

The supported language set is the following: English, Spanish, Italian, German and Dutch.

Besides the individual components, guidelines exists on how to add languages and how to adjust components for specific situations and topics.

In the Figure 3.8, there is the main schema of the work flow in OpeNER. This work flow could be seen as a processing chain for each language in OpeNER. The input regarding this schema is prepared to be raw text and the output of all the modules will be KAF [34].

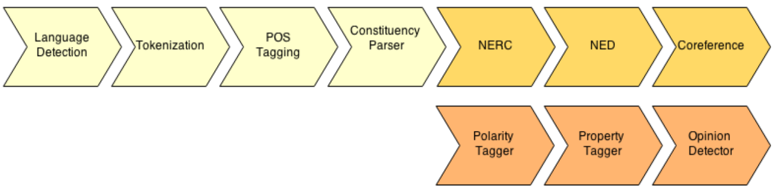

Figure 3.8: OpeNER Schema

Next we describe each module of the opener work flow.

Language Detection

This component is the responsible of detecting the language of an input document and delivers it to the correct language pipeline. If the language detector detects that a text is in English the input text should send the text flow to the English pipeline and so forth.

Tokenization

This component is the responsible of tokenizing the text in two levels, in sentence level and in word level. This component is crucial to the correct working of the rest of NLP components. This component is the first one of each language processing chain.

Part of Speech Tagging

This component is the responsible of assigning each token its morphological label. This component is crucial to the correct working of the rest of NLP components. This component also includes the lemmatization of the words.

Constituent parser

Parsing means providing the syntactic tree representation of a sentence. The component provides shift-reduced style constituent parsers for English, French, Italian and Spanish trained using the Apache OpenNLP API. For Dutch and German Alpino and Stanford parsers are required respectively. Constituent parsing is primarily used in OpeNER as an input to the Coreference resolution system.

Named Entity Resolution

Named Entity Resolution consists of processing named entities in text. The overall objective is to be able to recognize, classify and link every mention of a specific named entity in a text. By Named Entity we usually mean a proper name of a person, a place, an organization, etc.

A named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities which make them ambiguous. For example, the form 'san juan' can be used to ambiguously refer to dozens of toponyms, persons, a saint, etc.

Named Entity Recognition and Classification

Generally, since MUC and CONLL [35] shared tasks, NERC uses manually annotated data which serves to train machine learning models in a supervised manner. More recent trends aim at building automatic silver-standard and gold-standard datasets from existing large knowledge resources such as Wikipedia [36] to avoid the reliance on hand-generated data for training.

NER taggers recognized a variety of Named Entity types, namely, references to PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS, although in principle, given the appropriated annotated data, any type of Named Entity can be recognized. In this sense, OpeNER will, during its second year of development, be looking at recognizing and classifying Named Entity types releated with the Tourist domain such as restaurants, hotels, and perhaps monuments, theatres, etc.

Named Entity Disambiguation (NED)

Named Entity Recognition and Classification (NERC) deals with the detection and identification of specific entities in running text. Once the named entities are recognised they can be identified or disambiguated with respect to an existing catalogue. This is required because the "surface form" of a Named Entity can actually refer to several real things in the world. Wikipedia has become the de facto standard as such a named entity catalogue. Thus, if the form 'San Juan' appears in a given document, the NED task consist of deciding to which of the "San Juan" things listed in Wikipedia is actually the "San Juan" source form in that document referring to [37].

In OpeNER the NED component is based on the DBpedia Spotlight [38] which uses the DBpedia as the catalogue to perform the disambiguation. Within the OpeNER project new NED tools have been developed for each of the languages based on the English DBpedia Spotlight.

Coreference Resolution

Coreference resolution aims at grouping together all the mentions in the text to a Named Entity. For example, a person can be referred to by using her proper name, Claudia Lawrence, or by other types of expressions, such as "her", "she", "the 35-year-old", "Peter Lawrence's daughter", "the university chef", etc. Coreference resolution will aim at linking together or "clustering" every mention to a given Named Entity. This is useful, for example, if we really want to know who is talking about what. In the Tourist Domain, it is hoped that the coreference resolution will help to clarify who says an opinion about which hotel.

Polarity tagging

The polarity tagging is a task by means which terms in a text are tagged with their correct polarity and sentiment modifier label.

Words with polarity are words that express a negative or positive opinion, belief, attitude, etc. towards a certain topic. In our case polarity refers to out of context or "prior" polarity, i.e. to words which evoke something positive or negative independent from the context in which they are found. Polarity words can be nouns, verbs, adjectives and adverbs.

Opinion detection

The opinion detection is concerned with the identification of opinions in a text at the expression level. This task has received a lot of interesest in last period because of the explosion of the social networks. More and more companies use social networks to promote and offer their products, and they receive a lot of feedback from their customers as well. Considering the thousands of reactions that the people generates every on Social Networks, automatic analysis techniques become more and more interesting for extracting automatically opinions from this data. This is not only about deciding if a text is in general expressing a positive or negative opinion, but detecting and extracting single opinions and the entities that build these opinions:

1. Opinion expression: expressions that indicate emotions, sentiments, opinions or other private states
2. Opinion holder: mentions of whom is the opinion from
3. Opinion target: expressions that indicate what the opinion is about

For instance consider the sentence "I like the design of Ipod video". These are the elements of the opinion extracted from it:

1. Opinion expression: like
2. Opinion holder: I
3. Opinion target: the design of Ipod video

3.5 Visualization tool

3.5.1 Django

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, and we can focus on writing our app. It's free and open source.

The programming paradigm is based on the architecture pattern Model-View-Template, as we can see in Figure 3.9.

Model is the data access layer. This layer contains anything and everything about the data: how to access it, how to validate it, which behaviors it has and the relationships between the data. Template is the presentation layer. This layer contains presentation-related decisions: how something should be displayed on a Web page or other type of document. View is the business logic layer. This layer contains the logic that accesses the model and defers to the appropriate templates. You can think of it as the bridge between models and templates. URL Dispatcher is the bridge between the views and the browser, its responsibility is to call to the view according the URL mapped.

We use the django web framework for develop the web application for show the results.

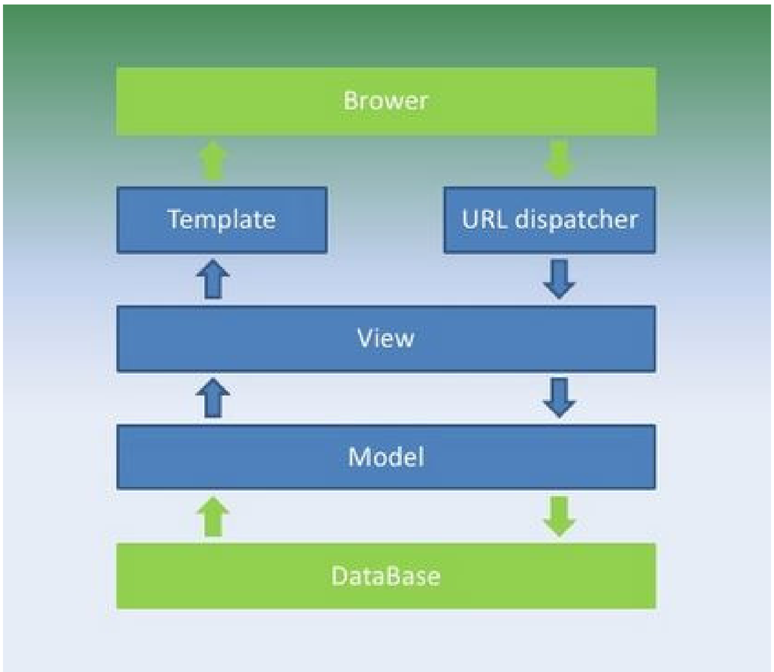

Figure 3.9: Django Framework Architecture

Chapter 4

Development

This chapter introduces the development steps of the two strategies presented in this project: Tweets Analysis Strategy (TAS), described in Section 4.3.2 and Cluster Analysis Strategy (CAS), described in Section 4.4.2. All the development of this project has been done in a Hadoop cluster deployed in some virtual machines managed by Open Nebula [39].

4.1 Strategies

In order to address the problem of making a prediction of the political elections outcomes, we have designed two strategies based on analyzing user's profiles for identifying their political tendency.

The aim of the first one, named Tweets Analysis Strategy TAS, is to analyze users tweets, identifying its polarity and the political party referred. Once, we've analyzed it, we calculate the trends of the users and the result is the prediction of their political tendency.

The second one, named Cluster Analysis Strategy CAS, is based on clustering users groups by their friendship in Twitter. This is achieved by modeling users in a graph and applying the Label Propagation Algorithm [40]. Then, we've identified users groups with strong relationship between them. Next, we relate and look for the maximum matching between political parties and the user groups. The final result of the algorithm returns the number of twitter users related to each political party.

4.2 Definition

This section describes the general definitions and notations used throughout this chapter. First, we need to distinguish between different kind of tweets and Twitter users, which will play a role in a particular political context. Let T be the set of all the tweets and U be the set of all the twitter users,

$$T = \{t | t \text{ is a Twitter Tweet } \},$$

$$U = \{u | u \text{ is a Twitter User } \}.$$

Let $C \subseteq T$ be the Political Tweets Set as the subset of tweets to be analyzed. In our proposal, they are those tweets analyzed which talk about Spanish political issues.

Let $O \subseteq U$ be the Political Users Set as the subset of users who launch some tweet of the C set. In our proposal, they are some users who talk about Spanish political issues.

Let $X = \{1, ..., n\}$ be the set of Political Parties in Spain.

Let P, $P \subseteq U$, be the Politicians set. Its users are Spanish politicians or recognized sympathizers clearly identified a priori. Let $P_x = \{p \in P | p \text{ sympathizes with the political party } x\}$. Notice that $\{P_x\}_{x \in X}$ is a partition of P

For instance, in our case, $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$ where $1 = PP, 2 = PSOE, 3 = Ciudadanos, 4 = Podemos, 5 = PDeCAT, 6 = CUP, 7 = ERC, 8 = IU$.

An example of an user of P_1 can be *@marianorajoy* that is the twitter username of the president of the Partido Popular or *@Sorayapp* Member of the congress of the deputies and deputy of Partido Popular.

4.3 Tweets Analysis Strategy (TAS)

4.3.1 Specific definitions

In order to describe the TAS strategy we need to fix some specific definitions and notations.

A *hashtag* is a word that starts with "#", that might be found within in a tweet text. For instance: *#pp, #ppostureo,...*

At the beginning of the TAS strategy, we define the *Keywords Set* denoted by K. This set contains some Political Users from O, some Hashtags with political content and some political-related words. For instance: *@marianorajoy, @Sorayapp, #pp, #psoe, podemitas, pperos,...*

To classify the tweets, we will consider C_x as the *Collection Keywords Set related to the Political Party* x, for a given Political Party x. This set contains the identifiers of Political Users from O, some Hashtags with political content and some political-related words, all of them related to the Political Party x.

Notice that $K \subseteq \bigcup_{x \in X} C_x$.

Let B, $B \subseteq T$, be the set of Negative Tweets. It is a set of tweets with a negative sentiment; it means that contains negative emoticons. For instance: *"La manzana podrida eres tu @marianorajoy :(", "La gestion del PSOE es nefasta :("*

Let G, $G \subseteq T$, be the set of Positive Tweets. It is a set of tweets with a positive sentiment; it means that contains positive emoticons. For instance: *"Lo mejor que le puede pasar a España, es que gane el PP :)", "Yo voto #PSOE :)"*

4.3.2 TAS Strategy

The TAS strategy is based on getting the Political Tweets C, analyze their sentiment and classify each Tweet according to the Political Party x that it is related to. Each time a Tweet is analyzed we obtain in return its sentiment and the Political Parties related to it. Notice that some tweets may refer to more than one political party. When we have analyzed all the Tweets for one user, we can compute the overall polarity for each Political Party concerning this user. Hence we have the political tendency of a given user. The last step is based on computing all the tendencies of all users to obtain the global tendency.

The TAS strategy is designed in the following three main blocks, that are sketched in Figure 4.1:

Data Extraction The main goal is to extract Tweets from the Twitter API, which are related to political context. See more explanation in 4.3.4.

Sentiment Analysis This block is focused in analyzing the sentiment of a Tweet. See more explanation in 4.3.5.

Data Classification This process assigns a Tweet to its related political parties. See more explanation in 4.3.6.

The input to the Data extraction module is a tweet and the output of the global of the process is an analyzed tweet, which is composed by the tweet text itself, the sentiment of the tweet and the political parties related to the tweet.

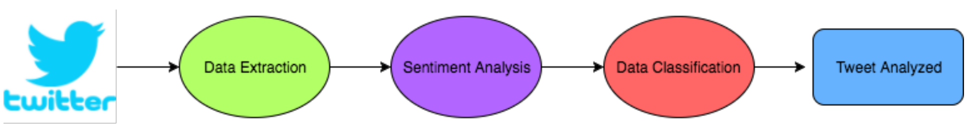

Figure 4.1: Tweet Analysis Lite Architecture

4.3.3 Topology

In the next Figure 4.2, we can see an expanded topology based on the previous Figure 4.1. From Figure 4.2, we can see the topology to extract and manage the tweets in real time, powered by Storm [24] where we can distinguish the three main blocks: Data Extraction, Sentiment Analysis and Data Classification. Inside of the blocks, we can appreciate the related technologies used in each process and the main operations described in the next points.

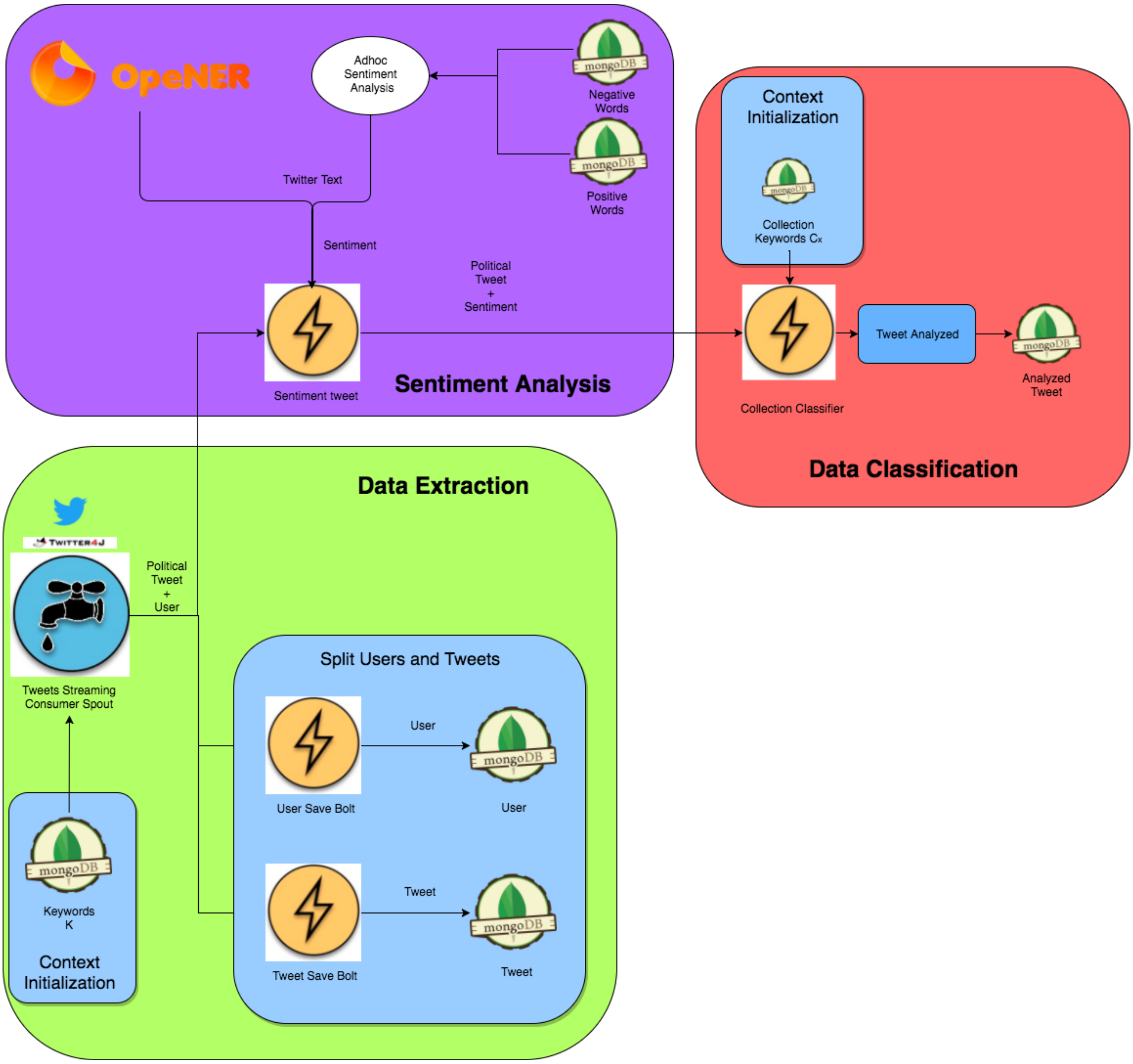

Figure 4.2: Architecture Tweet Analysis Strategy

Next sections describe in detail each of the previous blocks.

4.3.4 Data extraction

The data extraction process is devoted to capture the tweets in C. This process relies on the Keywords set K. Besides, it is necessary to clean the data for the next analysis. In the next figure 4.3, we can see the spout and bolts introduced in the next points.

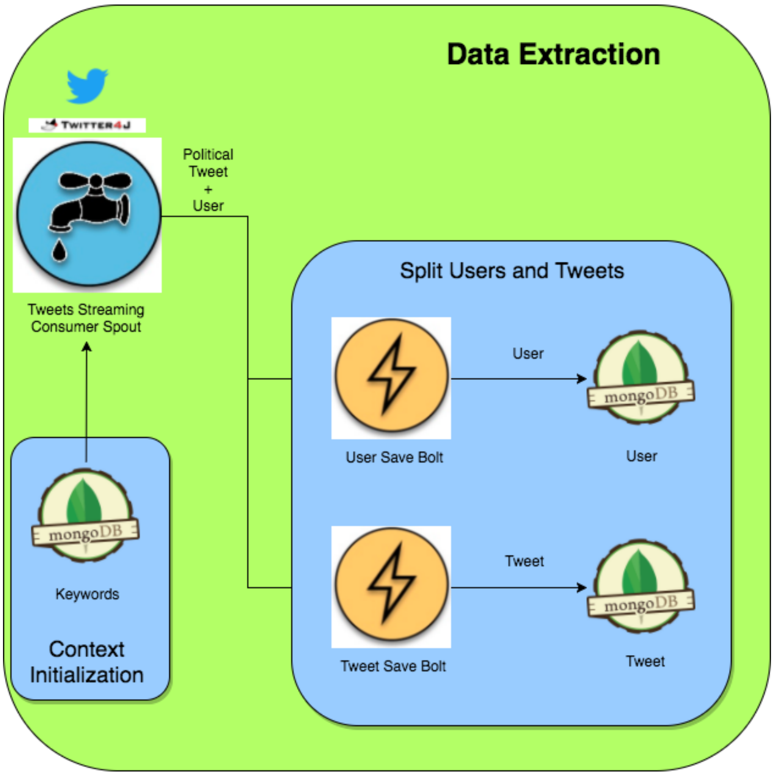

Figure 4.3: Architecture Tweet Data Extraction

Context initialization

When performing a tweet analysis, we need to focus on a particular target context. For instance, in our case, we are interested in the Spanish political context. Other examples could be analyzing the tweets referred to a particular sport competition, or most popular items.

To initialize the context of the project we required two data sets to contextualize the analysis. These sets have to be defined before running the project. The first is the Keywords set K that contains the keywords used to extract the Political Tweets. Hence, those tweets containing any of the words in K will be captured. Some example of the Keywords set K is shown in the Figure 4.4, where we can see user entities starting with @, hashtags starting with # and political words like ”C’s”.

@PPopular
@Pablo_Iglesias_
@GirautaOficial

#Podemos
#ConIlusion
#AlbertPresidente

PSOE
Rajoy
C's

Figure 4.4: Example of some words in K Keyword for Tweet Extraction in the Spanish political context

Once we have fetched the target tweets, we will proceed to classify them. For such purpose, we use a second set of words, Collections Keywords set C_x that contains the Keywords to classify the Tweets according to each political party x. Some example of the Keywords Collection set is shown in Figure 4.5 where we can see an example of some keywords used to classify the Political parties of PSOE and Partido Popular.

Collection Keywords

C_1 Partido Popular	C_2 PSOE
partidopopular	PSOE
pp	#FemForaRajoy
#EspañaEnSerio	#SomLaSolucio
#PP	#PSOE
#YoVotoPP	#OrgulloSocialista
@marianorajoy	#VOTAPSOE
@AlfonsoAlonsoPP	#PedroPresidente
@PPopular	@sanchezcastejon
@Sorayapp	@PSOE
@mdcospedal	@miqueliceta

Figure 4.5: Example C_1 = Partido Popular and C_2 = PSOE Collection Keywords

The keywords of the C_x Collection Keywords can be similar to K Keywords Set but not necessary equal. This is because some words that appear in the C_x set could produce a confusion during the extracting process of the tweets. For instance the word ”podemos” could get tweets which are not in the political context. However, they might be useful in the classification process.

In this project all of the keywords of the K set and C_x set were found in the X Political Parties twitter account and web pages. In sections 7.1 and 7.2 of the appendix we show the keyword collections used for this project.

Tweets Streaming Consumer Spout

For the tweets extraction, we use the Twitter API [41], specifically the StreamingAPI [31], because it provides real-time tweets and user information. For getting access to the API, we must create a developer account with a Twitter account and it gives us the credentials for calling the client API.

The implementation of the client is allowed in the Tweets Streaming Consumer Spout. Storm defines a Spout like a streaming data input of the project. In order to provide the Twitter client API in the Storm project, we use the twitter4J [26] library, which provides all the calls of the twitter API. In order to extract the context tweets, we give the keywords to twitter4J in the filter parameter named ”track”. To obtain the political tweets C, we use the track function of the Twitter API. It allows us to obtain tweets that match with any of the words in the Keywords set K. When we capture a tweet, the spout sends to the linked Bolts. We can find an example of what the streaming API returns in the Appendices 7.4.

With this strategy, we have found some problems with ambiguous Keywords named ”soft keywords” concretely that can't be neither a Hashtag nor a user, because those words could have different meanings, so the procedure could return non political tweets. An example of a soft keyword could be ”podemos”, since it can be the political party or the conjugation of the verb ”poder”. To solve this problem, we have identified the soft keywords and when it appears in a tweet, we search for other matching with another non-ambiguous word. If we find a soft keyword with another strong keyword, we analyze the tweet. On the other hand, if the soft keyword is alone then we will discard the tweet.

It is worth pointing out that we use the filtering option of the API to get tweets written in Spanish.

Split users and tweets

The Twitter API returns information about both the User and Tweet. To reduce the database space and improving the post-processing data, we separate the User and the Tweet information before saving in the *mongodb* database. This action is done in the *Tweet Save Bolt* and *User Save Bolt*.

4.3.5 Sentiment analysis

The objective of the Sentiment Analysis block is to get a positive, negative or neutral value per tweet text. To obtain it, we searched different tools to analyze the tweets.

Likewise, for the sake of enriching the information of the tweet text, we extract the content of the URL appeared in the tweet text, when it is applicable. When the content is a news headline, we get the title and if the content is a tweet, we get the complete tweet text. This is done because we detected that many of the tweets had some URLs in the text and we lost information about the meaning of the text if we excluded the URL content.

At the beginning, we tested and deployed the Opener project [42]. This is a project funded by the European Commission and it aims to be able to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts. This project is composed by different API calls to get the sentiment of a text. Unfortunately, the API calls could not be deployed correctly in our local server. Therefore, we needed to call external API. As a consequence the execution time was increased drastically. Furthermore, in the context of the tweet text, we have an extra difficulty: we can realize that many tweets are misspelled, so it adds more difficulty to the analysis process. For these reasons, we refused the usage of this system and we looked for another system that could fit better in the context of the tweet texts. Figure 4.6 shows the topology and the two ways explored to analyze the sentiment of the tweets: *Opener project* and *Own sentiment Analysis* thats we will introduce in the next paragraph.

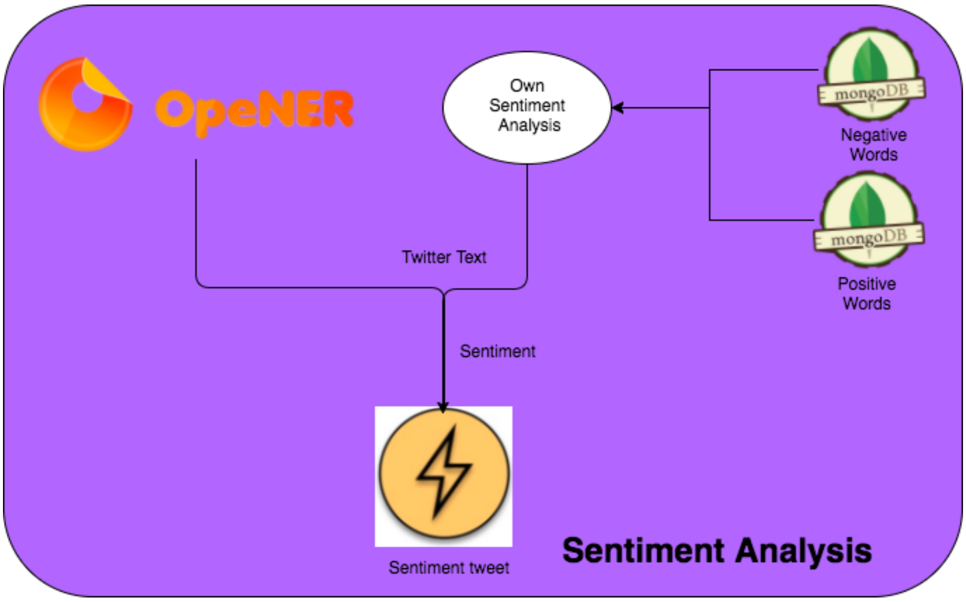

Figure 4.6: Architecture of the Sentiment Analysis

As an alternative to Opener, we decided to implement an adhoc sentiment analysis system. In order to get the sentiment of a tweet, we match and count the positive and negative words and emoticons inside of a tweet. The final result between the sentiment analysis is the *difference of* the positive and negative words/emoticons, which can be positive ($difference > 1$), negative ($difference < 0$) or neutral ($difference == 0$).

To perform this computation for each tweet, we need previous databases with positive and negative words. To build this database we have extracted 50,000 positive and negative tweets. As in [43], the positive or negative character of a tweet is inherited from its emoticons. The emoticons have been classified as follows:

Positive	Negative
:)	:(
:)	:-(
;-)	=(
:-)	: (
=)	
:D	

Figure 4.7: Positive and Negative Emoticons

Following this process the 100,000 tweets are classified in positive and negative tweets (Figure 4.8 shows how the twitter positive and negative spouts classifies the tweets into two different databases).

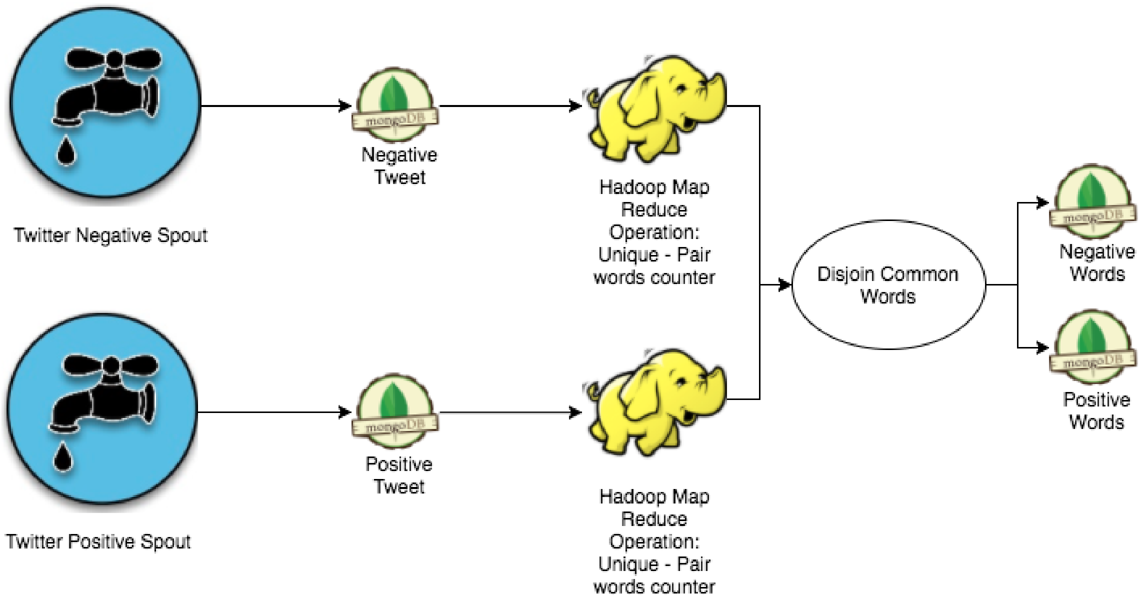

Figure 4.8: Architecture of the Own Sentiment Analysis

From these Tweet databases we will elaborate two words databases: one with positive words and the other with negative words. Previous to the word extraction process, we clean the tweets to discard non significant words. For instance we exclude the words that contains the symbols: #, @, http:// or https// and the words that contains "RT" (this is the acronym that appears in a retweet) and the words with less than three letters.

Then, in the word extraction process, for each tweet we elaborate a list of significant words of the tweet, and also a list with all the pairs of consecutive significant words. For instance in the tweet "*La manzana podrida eres tu @marianorajoy :(*", the list of significant words is: *manzana, podrida, eres*; and the list of significant pair is: *manzana podrida, podrida eres*. In Figure 4.9, we can see an example of the words extraction from three different tweets. In this example, since the emoticons that appears are negative, the extracted words will be collected in the negative set of words.

These positive and negative words are extracted using the Map Reduce operation of Hadoop [15].

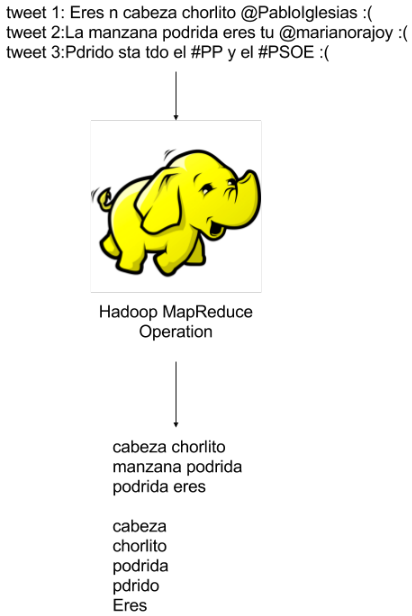

Figure 4.9: Example Map Reduce Own Sentiment

After this process, we have a list of positive and negative words/pairs of words. Notice that there may be common words in both list. These words have to be removed because they can be considered neutral words. The result of the overall process is a positive and negative collection words/pairs of words.

4.3.6 Data classification

For each tweet to be analyzed we want to classify it according to its political affinity. To do so we use the Collection Keyword set C_x, for each different political party x. Then we look for matching keywords within the Tweet and the words of every C_x. Notice that one tweet can be classified as affine to one or more Political Parties because we can find two keywords of different Political Parties. The final result of this process is a tweet together with the Political Parties related to it and the sentiment (positive, negative or neutral) of the tweet. All of this is done in the process *Collection Classifier*.

For instance, consider the tweet ”el @PSOE siempre favorece al @PP ”. Comparing it with the different C_x, we obtain his affinity to C_1 = Partido Popular and C_2 = PSOE. Counting the positive and negative words we have a positive sentiment, because ”favorece” and ”siempre” belongs to the positive words set.

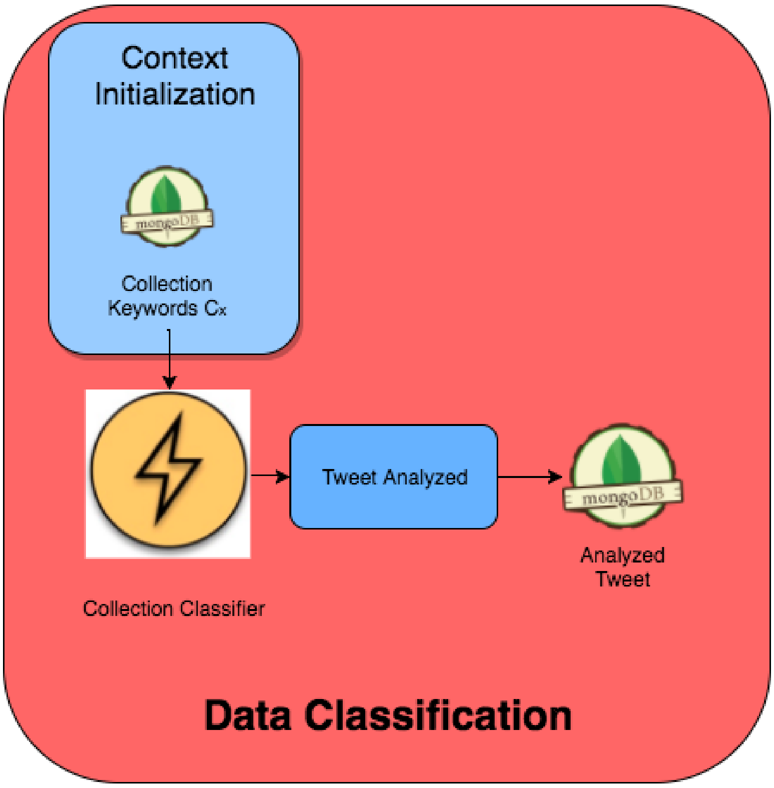

Figure 4.10: Architecture Tweet Classification Users

In the next Figure 4.11 we can see a tweet classification example and a Tweet Analyzed with its sentiment and the political parties related. In this case, the sentiment of the tweet is negative and it is related to the PP and PSOE parties.

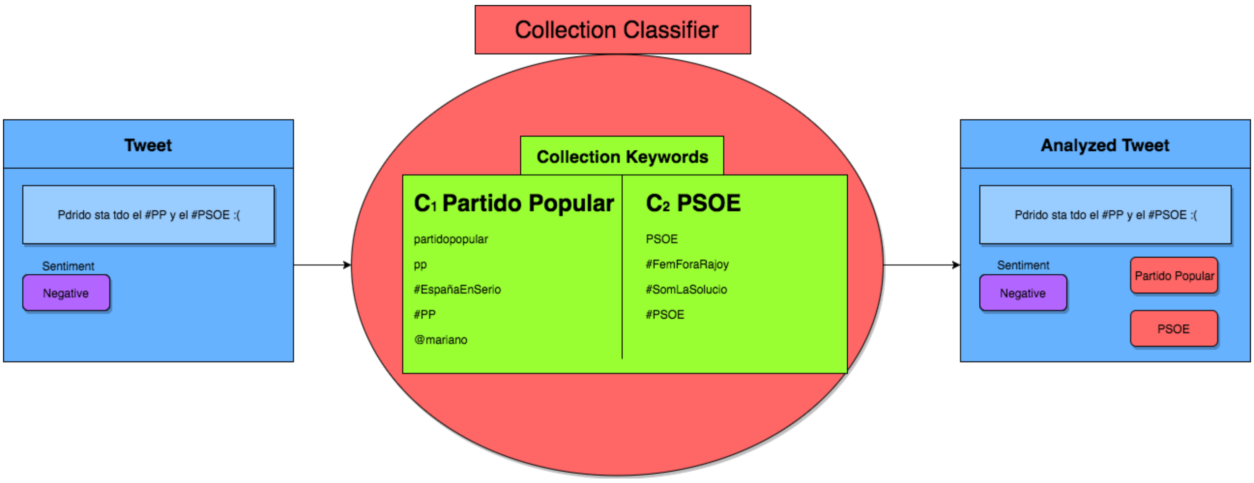

Figure 4.11: Example Analyzed Tweet

4.3.7 TAS Result

The final objective of the TAS strategy is to get the global political tendency. For such a purpose we need to analyze the tendency of each user. Hence we need to get all the analyzed tweets for each user and identify the political party occurrences and its sentiment.

Let n be the number of political parties, and let u be an analyzed twitter user who has launched m analyzed tweets. Then, we arrange the data by means of an $m \times n$ matrix $M_u = (c_{ix})$, where c_{ix} is the coefficient refereed to the tweet $t_i \in C$ launched by u and the political party $x \in X$.

The content of each cell c_{ix} of the matrix can take four possible values:

$$c_{ix} = \begin{cases} 1, & \text{if } t_i \text{ has positive sentiment and mentions } x, \\ -1, & \text{if } t_i \text{ has negative sentiment and mentions } x, \\ 0, & \text{if } t_i \text{ has neutral sentiment and mentions } x, \\ \text{NULL}, & \text{if } t_i \text{ does not mention } x, \end{cases}$$

One example can be shown in Table 4.1.

Mu	Political Party 1	Political Party 2	Political Party 3	Political Party 4	...	Political Party n
t1	1	1	1	1	...	...
t2	-1	NULL	-1	NULL	...	...
t3	NULL	NULL	NULL	NULL	...	...
t4	1	1	1	1	...	...
t5	-1	0	0	-1	...	...
t6	-1	1	0	1	...	...
t7	-1	NULL	-1	NULL	...	...
.	...	...	...	...	...	...
.	...	...	...	...	...	...
Tm	...	...	...	...	...	...

Table 4.1: Example of a weight matrix M_u for a given Twitter user u

Once, we have the data collected in M_u, we compute w_x as the global weight sentiment of user u related to each political party x, as follows:

$$w_x = \sum_{i=1}^{m} c_{ix}, \forall c_{ix} \neq NULL.$$

On the other hand, let r_x be the number of citations of political party x from all the tweets that u has launched, that is

$$r_x = \#\{c_{ix} : \ 1 \leq i \leq m, \ c_{ix} \neq NULL\}.$$

The political affinity s_x for a specific user u and for each political party x, is calculated as the product of the global weight sentiment w_x and the number of appearances r_x, as follows

$$s_x = r_x \cdot w_x.$$

Then the political tendency for user u is that political party x whose political affinity s_x attains the maximum value. Likewise, we have obtained the secondary preferred political party for user u according to the second maximum value of s_x.

In order to clarify this process, we show a simple example in the Table 4.2. We assume 6 political parties ($n = 6$) and 10 tweets ($m = 10$) for a given user u. The three last rows of the table show the values w_x, r_x and s_x for each political party. Notice that, in this example, the political tendency of this user is the political party $x = 4$, while his secondary tendency is political party $x = 5$.

Mu	Political Party 1	Political Party 2	Political Party 3	Political Party 4	Political Party 5	Political Party 6
t1	1	1	1	1	1	1
t2	-1	NULL	-1	NULL	1	NULL
t3	NULL	NULL	NULL	NULL	-1	NULL
t4	1	1	1	1	NULL	1
t5	-1	-1	NULL	-1	NULL	-1
t6	-1	1	NULL	1	NULL	1
t7	-1	NULL	-1	NULL	NULL	NULL
t8	-1	-1	NULL	1	NULL	-1
t9	NULL	-1	NULL	1	NULL	-1
t10	NULL	1	1	1	1	1
w_x	-3	1	1	5	2	1
r_x	7	7	5	7	4	7
s_x	-21	7	5	35	8	7

Table 4.2: Calculation example of a weight matrix M_u for a given Twitter user u

The global political tendency prediction is a chart which shows the number of users affine to each political party, according to the political tendency of each user computed above (see Figure 4.12).

The calculation of the global tendency is done in the Django web application, that gives us a graphical representation of the users global tendency and single user political tendency. In addition, the web platform provides information of the execution process and a management of the project. In the Figure 4.12 we can see an example of the Django application with the Global results, in this case only shows a sample of users analyzed.

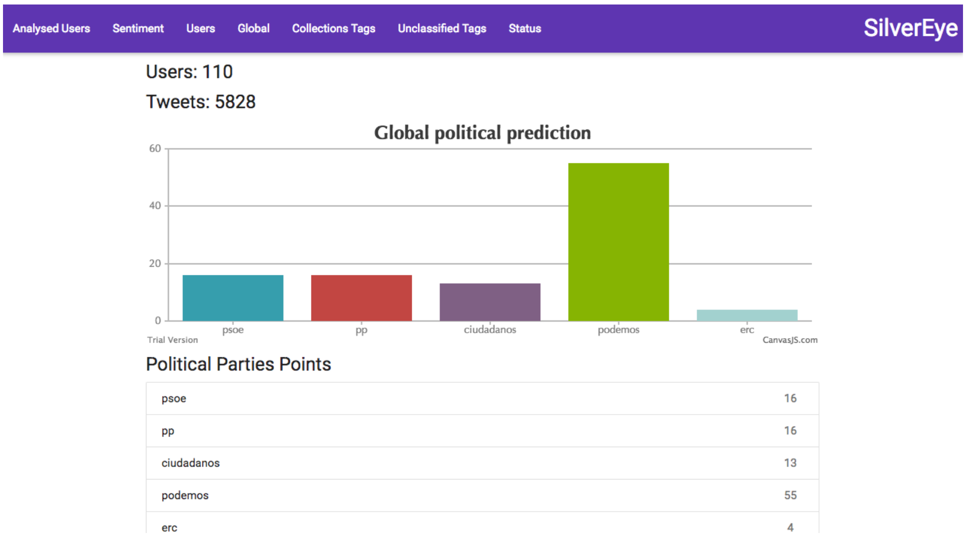

Figure 4.12: Example Analyzed Tweet

4.4 Clusters Analysis Strategy (CAS)

This second strategy differs from the previous since it is not focused on analyzing the opinions of users, but exploiting their affinity to other users. In this section we will highlight the main aspects of this strategy.

4.4.1 Specific Definitions

The CAS strategy exploits the information provided by the friendship relationship between twitter users. Such relationships can be modeled with graphs.

Definition 4.4.1. Let $D = (V, E)$ be the Twitter Affinity Digraph, where V is a set of twitter users, and $(u, v) \in E \Leftrightarrow u$ is following v

Definition 4.4.2. Let u and v be two twitter users. It is said that:

$$(u,v) \in E \Leftrightarrow v \text{ is friend of } u,$$

$$(u,v) \in E \Leftrightarrow u \text{ is follower of } v.$$

Then for any user $u \in V$ we define the friendship set as,

$$F_u = \{v \in V : (u,v) \in E\}.$$

Analogously the followers list is,

$$G_u = \{v \in V : (v,u) \in E\}.$$

For example, consider the digraph in Figure 4.13.

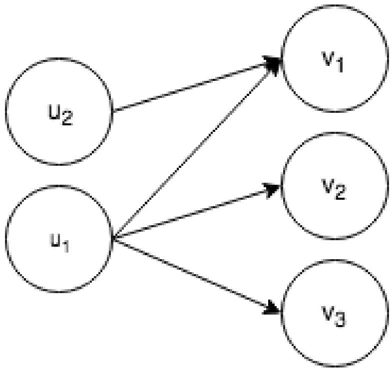

Figure 4.13: Example of a Twitter Affinity Digraph

In this digraph the friendship and followers list are:

$$F_{u_1} = \{v_1, v_2, v_3\}, G_{v_1} = \{u_1, u_2\}.$$

Definition 4.4.3. Consider a set $A \subseteq V$. Then the friendship set of A is

$$F_A = \bigcup_{a \in A} F_a.$$

Recall that O is the Political users set (that is users who launch political tweets in Spanish context) and P is the politicians set (the users identified a priori as sympathizers to a political party). Then we can consider the following friendship set

$$F = F_O \cup F_P$$

.

For example, consider the digraph in Figure 4.14:

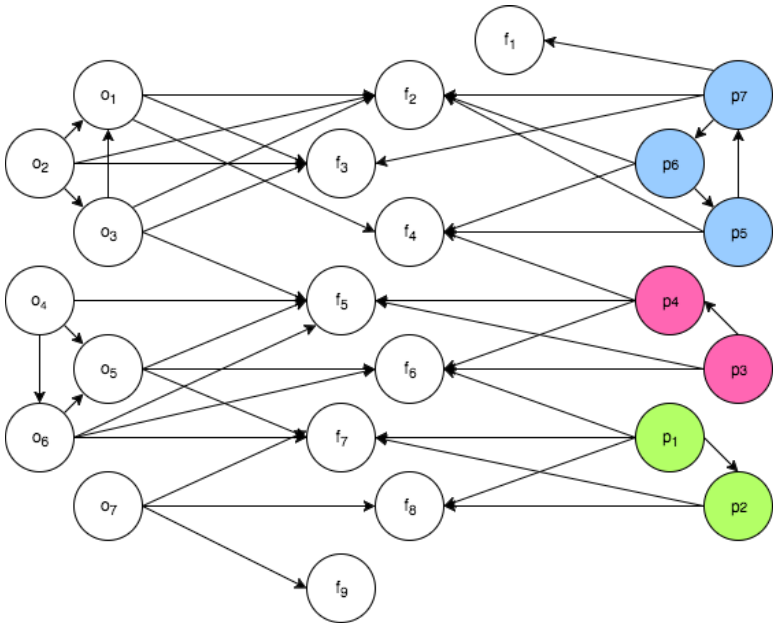

Figure 4.14: Users Digraph Example

Consider that there exist three political parties

$$X = \{1,2,3\}$$

We have a priori identified the respective lists of significant sympathizers:

$$P_1 = \{p_1, p_2\}$$

$$P_2 = \{p_3, p_4\}$$

$$P_3 = \{p_5, p_6, p_7\}$$

In the example digraph, the respective nodes are colored according to each political party.

Consider that our Object users set is:

$$O = \{o_1, o_2, ...o_7\}$$

Then, the friendship sets we can consider are:

$$F_O = \{f_2, f_3, f_4, f_5, f_6, f_7, f_8\}$$

$$F_P = \{f_2, f_3, f_4, f_5, f_6, f_7, f_8\}$$

$$F = F_O \cup F_P$$

Definition 4.4.4. Let the $D = (V,E)$ be the Twitter Affinity Digraph. Then we can consider the following Subdigraph:

$$D_O = (U,E) \text{ such that } U = O \cup F_O.$$

Figure 4.15 shows the subdigraph D_O in the previous example, which contains the objective users and their friends.

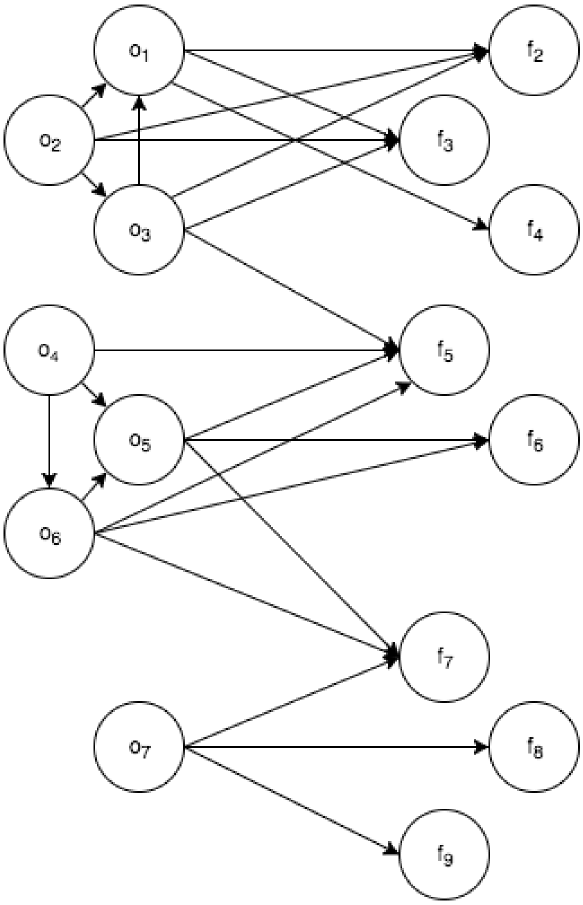

Figure 4.15: Example subdigraph D_O

4.4.2 CAS Strategy

This strategy is based on clustering their users according to the Twitter friendship. We have the assumption that the relation between the circle of friends in twitter (restricted to users who launch political tweets) is related to their political trends. With this assumption we design a strategy based on modeling the users O and her friends F_O in a graph.

When having the friendship users F_O and the Political Users O we join them in the $L = O \cup F_O$. Then we load the set L into a digraph using GraphFrames [22] and clustering the users according these friendships with the Label propagation Algorithm [40]. The output of this algorithm assigns one identifier of a group for each user. Likewise we can obtain a partition of the set L. Since our objective is to obtain Politician Users groups, we remove the F_O nodes. Hence the result is a partition of O, like $O = O_1 \cup O_2 \cup O_3 \cup \ldots \cup O_m$ where m could be the max number of users in the set V.

Likewise, we get the friendship users F_{P_x} of the Political Users. We will compare these sets with F_{O_i} to the detect the maximum coincidence and, hence, assign a political party to O_i.

In the next subsections we describe in detail all the strategy.

4.4.3 Topology

In Figure 4.16 we can see the topology of the strategy based on the extraction of the Political Users friends, the modeling of the digraph, the clustering of the users, the matching between the cluster O_i and F_{Px}.

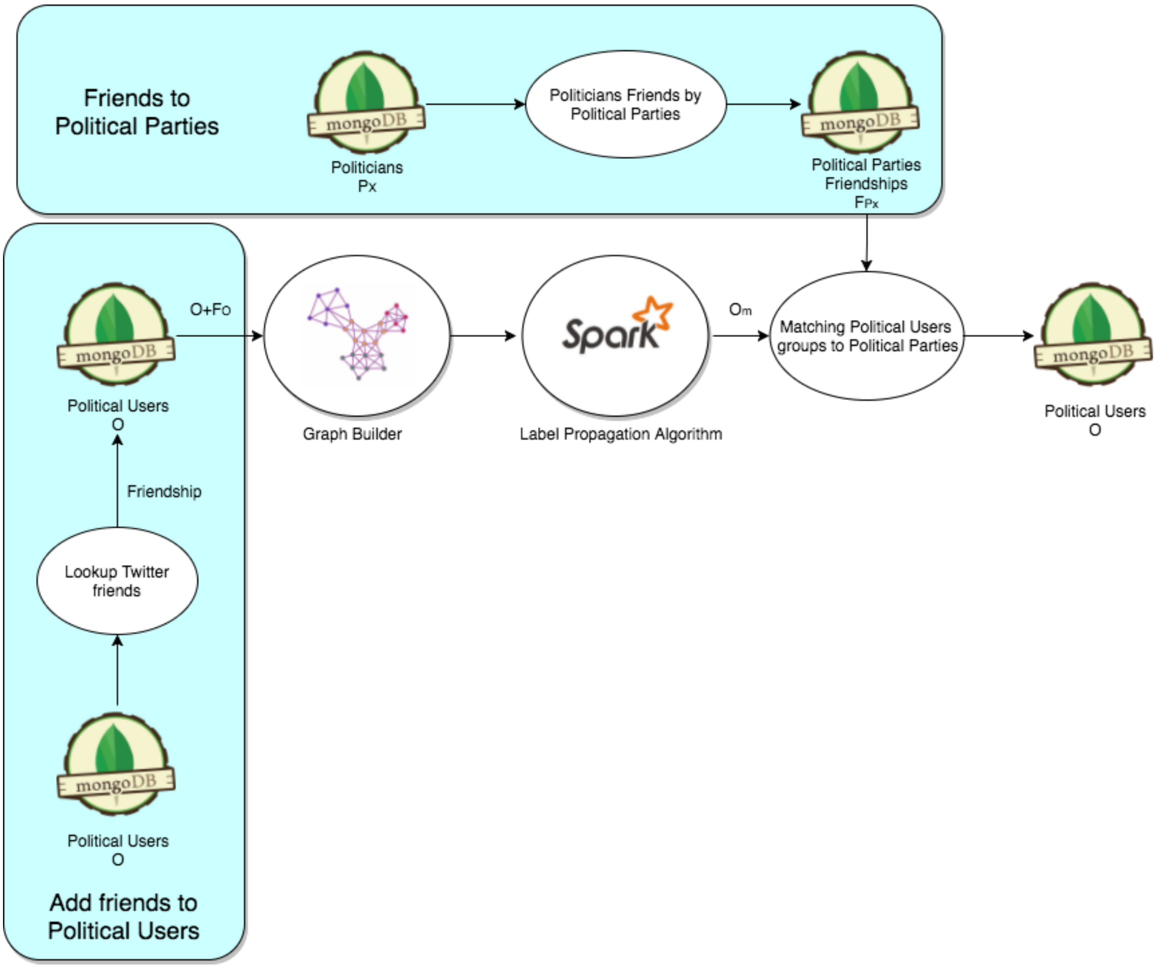

Figure 4.16: Architecture Clustering Users

4.4.4 Add friends to Political Users

To obtain the Political Users Set O, we proceed similarly as in TAS to fetch the tweets with political keywords 4.3.4. Then, using the Lookup twitter user API call [44] we can obtain the friendship set of $o \in O$, namely F_o, for each user o in O. All of them are stored in MongoDB database. Therefore MongoDB contains the Political Users with each friendship set.

4.4.5 Graph Builder

At this step, we model the objective users O and her friends F_O to a graph L, where the users are nodes and the friends are the edges D_V , for doing this we load the data to Spark [18] using the GraphFrames [22] library. The users are grouped in sets according to his friends by means of the Label Propagation Algorithm [40], which works as follows .

Suppose that a node x has neighbors $x1, x2, ..., x_k$ and that each neighbor carries a label denoting the community to which they belong to. Then x determines its community based on the labels of its neighbors. We assume that each node in the network chooses to join the community to which the maximum number of its neighbors belong to, with ties broken uniformly randomly. We initialize every node with unique labels and let the labels propagate through the network. As the labels propagate, densely connected groups of nodes quickly reach a consensus on a unique label. When many such (consensus) groups are created throughout the network, they continue expanding outwards while it is possible to do so. At the end of the propagation process, nodes having the same labels are grouped together as one community.

We perform this process iteratively, where at every step, each node updates its label based on the labels of its neighbors. The updating process can either be synchronous or asynchronous. In synchronous updating, node x at $t^t h$ iteration updates its label based on the labels of its neighbors at iteration $t-1$. Hence, $C_x(t) = f(C_{x1}(t-1),...,C_{x_k}(t-1))$, where $c_x(t)$ is the label of node x at time t. Hence we use asynchronous updating where,

$$C_x(t) = f(C_{x_1}(t),...,C_{x_im}(t),C_{x_i(m+1)}(t-1),...,C_{x_ik}(t-1)) \text{ and } x_{i1},...,x_{im}$$

are neighbors of x that have already been updated in the current iteration while $x_{i(m+1)},...,x_{ik}$ are neighbors that are not yet updated in the current iteration. The order in which all the n nodes in the network are updated at each iteration is chosen randomly Note that while we have n different labels at the beginning of the algorithm, the number of labels reduces over iterations, resulting in only as many unique labels as there are communities. Ideally the iterative process should continue until no node in the network changes its label. However, there could be nodes in the network that have an equal maximum number of neighbors in two or more communities. Since we break ties randomly among the possible candidates, the labels on such nodes could change over iterations even if the labels of their neighbors remain constant. Hence we perform the iterative process until every node in the network has a label to which the maximum number of its neighbors belong to.

By doing so we obtain a partition of the network into disjoint communities, where every node has at least as many neighbors within its community as it has with any other community. If $C_1,...,C_p$ are the labels that are currently active in the network and $d_i^{C_j}$ is the number of neighbors node i has with nodes of label C_j , then the algorithm is stopped when for every node i,

If i has label C_m then $d_i^{C_m} \geq d_i^{C_j} \forall j$

At the end of the iterative process, nodes with the same label are grouped together as communities. The stop criterion characterizing the obtained communities is similar (but not identical) to the definition of strong communities proposed by Radicchi et all [45]. While strong communities require each node to have strictly more neighbors within its community than outside, the communities obtained by the label propagation process require each node to have at least as many neighbors within its community as it has with each of the other communities. We can describe their proposed label propagation algorithm in the following steps.

The process has 5 steps:

1. Initialize the labels at all nodes in the network. For a given node $x, C_x(0) = x$.

2. Set $t = 1$.

3. Arrange the nodes in the network in a random order and set it to X.

4. For each $x \in X$ chosen in that specific order, let $C_x(t) = f(C_{x_{i1}}(t),...,C_{x_{im}}(t),C_{xi_{(m+1)}}(t-1),...,C_{x_{ik}}(t-1))$. f here returns the label occurring with the highest frequency among neighbors and ties are broken uniformly randomly.

5. If every node has a label that the maximum number of their neighbors have, then stop the algorithm. Else, set $t = t+1$ and go to (3).

The Label Propagation Algorithm is implemented in the GraphFrame library and allows to define the number of iterations to execute by the algorithm and we execute the algorithm m times with n iteration, where m and n can be $m = 1,2,3...10$ and $n = 1,2...10$. We define the value of m and n with the aim to find the iteration with the minimum number of groups. The algorithm ends whenever it returns a minimum number of groups equal to the the number of political groups x or when $m = 10$.

The result is a partition of the Political Users set O classified by different groups like $O = O_1 \cup O_2 \cup O_3 \cup O_m$, but not political groups, at the moment, only groups built by the friendship of these.

4.4.6 Political users

In order to relate the political parties with the partitions of objective users like $O_1, O_2, ..O_x$, we need to obtain the friendship of the Politician User P for each Political Party. Then we obtain the Politician friends using the Lookup twitter user API call [44] and consequently, we achieve the friendship set of $p \in P$, named F_p, for each user p in P. Therefore for each p in P_x we classify to a Political Party x and we assign the friendship for each Political Party and the result named F_{P_x}, for each Political Party x in X. On the whole we save the set F_X to a MongoDB named Political Friendships.

In Figure 4.17 we can see an example of the Politician friends where $P_1 = \{p_7, p_6, p_5\}$ and $P_2 = \{p_4, p_3\}$ and the $F_{P_1} = \{f_2, f_3, f_4\}$ and the $F_{P_2} = \{f_4, f_5, f_6\}$.

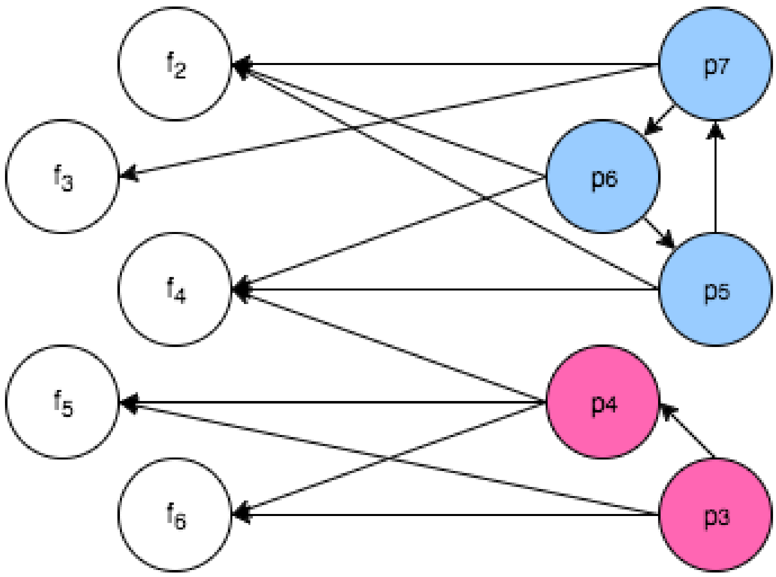

Figure 4.17: Political Friends

4.4.7 Matching Political Users groups to Political Parties

At this point, we have the Political Users O classified by partition like $O = O_1 \cup O_2 \cup O_3 \cup ... \cup O_m$, along with the set of friends of each group F_{O_i}. In the other side we have the Politicians set P and its partition $\{P_x\}_{x \in X}$. For each P_x we have their friends set F_{P_x} (See Figure 4.18).

In order to assign a proper political party x to each O_i, we count the maximum number of common nodes between F_{O_i} and every F_{P_x}. In Figure 4.18 we can appreciate an example that we introduce next: In one hand, we have the Political users O partitioned by $O_1 = \{o_1, o_2, o_3\}$, $O_2 = \{o_4, o_5, o_6\}$, $O_3 = \{o_7\}$. Related with their friends F_{O_i} where, $F_{O_1} = \{f_2, f_3, f_4, f_5, o_1, o_2, o_3\}$, $F_{O_2} = \{f_5, f_6, f_7, o_4, o_5, o_6\}$ and $F_{O_3} = \{f_7, f_8, f_9\}$. In the other hand, we have the politician users P partitioned by $P_1 = \{p_7, p_6, p_5\}$ that is related to the political party x_1, $P_2 = \{p_4, p_3\}$ that is related to the political party x_2 and $P_3 = \{p_1, p_2\}$ that is related to the political party x_3. Related with their friends F_{P_i} where, $F_{P_1} = \{f_1, f_2, f_3, f_4, p_5, p_6, p_7\}$, $F_{P_2} = \{f_4, f_5, f_6, p_4\}$ and $F_{P_3} = \{f_6, f_7, f_8, f_9, p_2\}$.

In this steps we will get the highest number of elements of the intersection for each F_{O_i} to for each F_{P_i}. For example, for the O_1 we have the friendship F_{O_1} and we will find the max elements with the intersection of F_{P_1}, F_{P_2} and F_{P_3}. The result is: $F_{O_1} \cap F_{P_1} = \{f_2, f_3, f_4\}, F_{O_1} \cap F_{P_2} = \{f_5\} and F_{O_1} \cap F_{P_3} = \emptyset$. Because $F_{O_1} \cap F_{P_1}$ have 3 elements we say that is the first political party of O_1 is x_1, and the second political party is x_2 because have one element in the intersection.

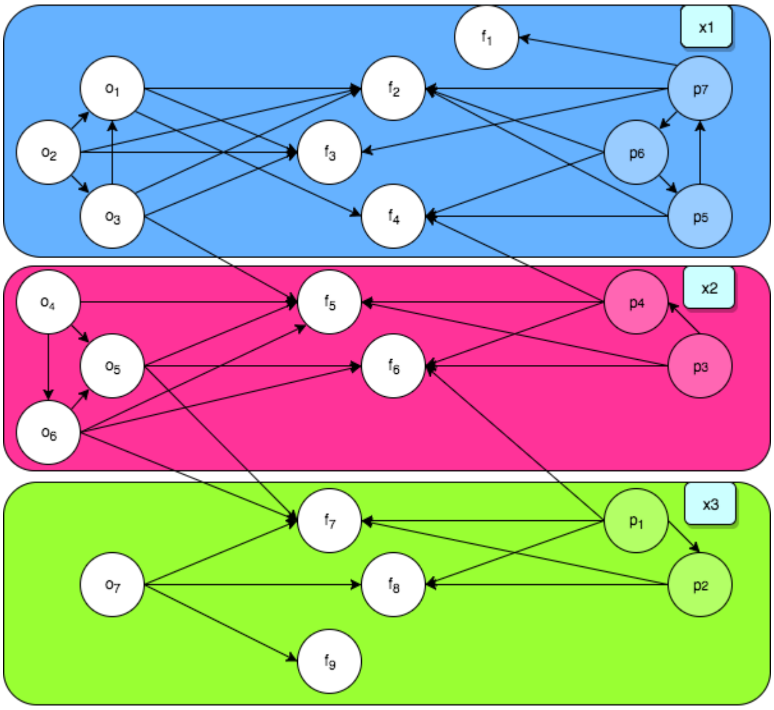

Figure 4.18: Relationship between clusters with friends

The last step is to assign the first and second political party x for each user.

4.4.8 CAS Result

The final result of the CAS strategy is the political tendency of a user, also with TAS strategy, CAS strategy gives us the first and second most probably political tendency of the user. The result is saved in Mongodb in the collection *AnalyzedUsers*.

Chapter 5

Test and Results

This chapter describes how we get the set of political users analyzed, which are used to compare with the results of the prediction. Likewise, the results of the TAS, CAS strategy and the sentiment analysis algorithm are shown and discussed.

5.1 Users Analyzed (TEN)

In this section with the aim of checking and validating the results, we will analyze manually the Tendency of the Users Analyzed set (TEN). In order to obtain TEN, we need a set of users whose political tendency has been manually analyzed. This is a challenge because the task to analyze manually a twitter user is not trivial given that common users post few political tweets and some of them contain irony, images or neutral text. The way for analyzing the profiles is to identify tweets supporters to a political party or tweets opposites to a political party. Also, when a user retweets a tweet sent by a politician user, we count this as a positive tendency to this political party. We have considered a set of 184 users, but we have reduced to 120 after discarding those accounts related to man media not Spanish users, prank accounts or TV programs.

According to the problems described above to analyze and identify the political tendency of a twitter user, we define the following criteria to analyze the trend of each user:

- Count the retweets of political representation of some party.
- Analyze the tweets image to search the positive or negative trend to some party. In this, we can find some politician images or images related to a political parties. These images give us some indicators of the political party orientation.
- Read the tweets.
- Read the users description.
- Analyze the profile and landing image.

The results are shown in the Annex 7.1. From the Table 7.1, we can see as for each row, we can find one user together with his positive or negative tendency for the different political parties analyzed. Apart from the problems described above, the complexity of this analysis is that users do not always emit political tweets, so the analysis of a user spends 5 minutes on average, because we need to analyze the major tweets of his profile. For this reason, it is not possible to analyze the preferences of some users in relation to all the political parties and as a consequence, those users have some positive or negative tendency empty. In those cases, we assume that the tendency can be either positive or negative. This list of users will be used throughout this chapter to compare with the results of the TAS strategy and CAS strategy.

For instance, some example of positive classification of one user to the political party "podemos" is with the retweet showed in Figure 5.1 of one politician of podemos. With the same user we also classify negatively to the "PP" political party given that the retweet 5.2 shows an image that talks about the corruption of the political party PP.

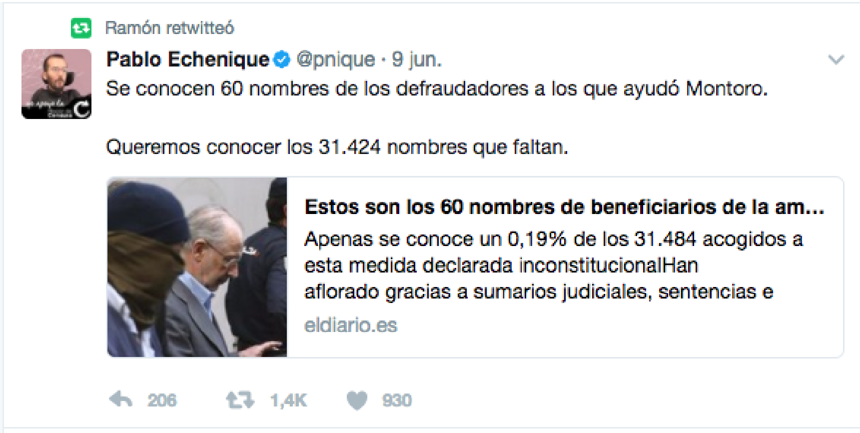

Figure 5.1: Retweet Example classified to the Political Party "Podemos"

Figure 5.2: Retweet classified negatively to the Political "Party PP"

5.2 Tweets Analysis Strategy (TAS)

5.2.1 Users analyzed

TAS strategy gives us the political tendency of the user based on the analysis of his political Tweets. The result of the TAS strategy is the main political tendency and the secondary tendency of the analyzed user. For checking this, we match the analyzed users (TEN) according with the following criteria.

- The political result of the algorithm matches with some of the positive parties identified to the user. In this case, we name it "*Succes*".
- The political result of the algorithm matches with some of the neutral parties identified to the user. In this case, we name it "*Neutral*".
- The political result of the algorithm matches with some of the negative parties identified to the user. In this case, we name it "*Error*".

5.2.2 Results

Table 5.1 shows the asserts, errors and neutral results for the main tendency given by the TAS strategy, which is named TAS1. The users have been classified in relation to the number of tweets sent by the user. Thus, the first row shows the results of the users, who have launched a range between 0 to 10 tweets/user, the second row users with a range between 0 to 25 tweets/user, the third row shows results with a range between 0 to 50 tweets/user and the last row shows the results related to users with a range between 0 to 100 tweets/user. We can see as the number of asserts increases drastically with the number of sent tweets from 26% up to 72,7%. In addition, Table 5.2 shows the same information related to the secondary tendency calculated by the TAS strategy named TAS2. In this case, we can see as the number of asserts decreases slightly when the numbers of tweets is increased. Specifically, it goes down poor 12% to 9%. In addition, as it was expected, we can see as the number of asserts is much lower than in the first tendency.

		TAS1					
Users	Tweets	#Assert	%Assert	#Error	%Error	#Neutral	%Neutral
97	10	26	26,80%	59	60,8%	12	12,37%
76	25	24	31,58%	42	55,3%	10	13,16%
56	50	22	39,29%	28	50,0%	6	10,71%
11	100	8	72,73%	2	18,2%	1	9,09%

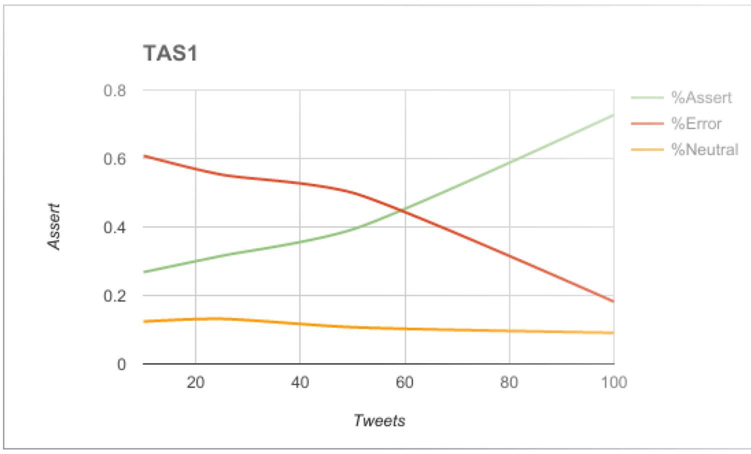

Table 5.1: TAS Main Results

		TAS2					
Users	Tweets	#Assert	%Assert	#Error	%Error	#Neutral	%Neutral
97	10	12	12,37%	72	74,2%	13	13,40%
76	25	8	10,53%	57	75,0%	11	14,47%
56	50	7	12,50%	42	75,0%	7	12,50%
11	100	1	9,09%	9	81,8%	1	9,09%

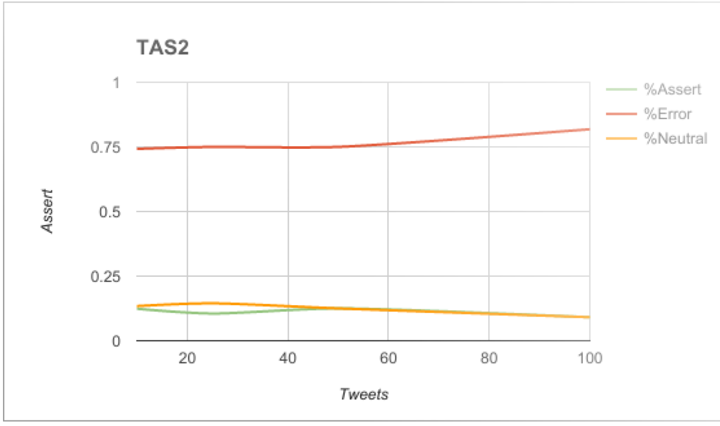

Table 5.2: TAS2 Results

5.3 Basic sentiment analysis results

5.3.1 Sentiment analyzed

The evaluation of the Tweets Sentiment Analysis is done getting a set of 183 tweets of TEN users. The first step is to analyze the tweets manually and assign a polarity for every tweet. The polarity can be positive (1), negative (-1) or neutral (0). Likewise, we execute our sentiment algorithm with the analyzed tweets. In the Annex 7.2 we can see the results obtained, where the "Real Sentiment" column has the sentiment analyzed manually, the "Result" column has the result of our own algorithm and the "Assert" column has the assert percentage.

5.3.2 Results

In order to analyze the result of our sentiment analysis algorithm, we consider three possible results in relation to the matching with the real sentiment of the tweet shown in Annex 7.2.

- When the real sentiment is positive or negative and the algorithm returns neutral we consider a 50% of assert.
- When the real sentiment is positive or negative and the algorithm matches correctly with the real sentiment, we consider a 100% of assert.
- When the real sentiment is positive or negative and the algorithm does not match correctly with the real sentiment we consider a 0% of assert.

Taking into account these considerations, the final results are shown in the Figure 5.3.

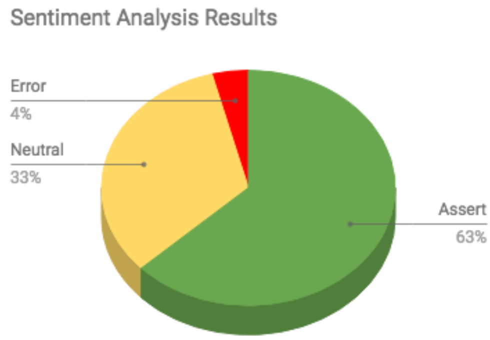

Figure 5.3: Sentiment Analysis Result

Likewise, if we remove the neutral tweets we have the results shown in Figure 5.4. In this case we can see as the percentage of asserts is increased up to 69%.

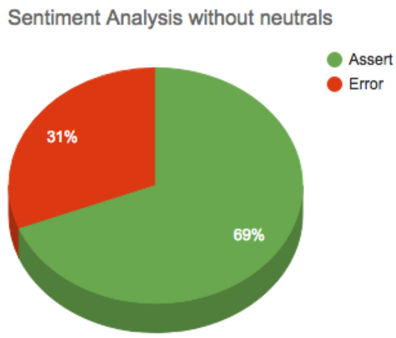

Figure 5.4: Sentiment Analysis Result

We can consider that these results are not bad given the simplicity of the algorithm used in this project. We wold like to point out that if we increase the number of positive and negative words, the number of neutral tweets would be decreased. However it is out of the scope of this project.

5.4 Clustering results (CAS)

5.4.1 Users analyzed

In this section, we will evaluate our sample and how we calculate and get the results. We execute the Cluster Analysis Strategy algorithm with the sample set of 120 users analyzed. The result of this analysis provides two different values. The primary political party and the secondary political party trend for each user. For checking the results we have only taken the first political party for each user. In Appendix 7.3 we can see the results obtained with the 120 users analyzed, described in Section 5.1. The comparison has been done as follows:

- The political result of the algorithm matches with some of the positive parties identified to the user, we named this "*Success*".

- The political result of the algorithm matches with some of the neutral parties identified to the user, we named this "*Partial Success*".
- The political result of the algorithm matches with some of the neutral parties identified to the user, but we have identified in the analysis one or more positive tendencies with some party, we named this "*Partial Success Error*".
- The political result of the algorithm matches with some of the negative parties identified to the user, we named this "*Error*".

5.4.2 Results

According to the previous considerations, we have the next results shown in Figure 5.5. As you can see, the success is 61,7% and the partial success is 16,7%. If we can consider assert the addition of the Success, Partial Success and Partial Error Success, then the total Assert is of the 99/120 (82,5%).

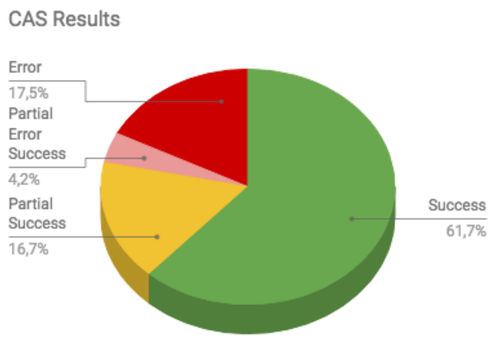

Figure 5.5: CAS Result

5.5 Political Tendency Results: TAS vs CAS Strategies

In this section we present the political tendency results, obtained in TAS and CAS Strategy, contrasted with the political tendency of the users analyzed (TEN) showed in Figure 5.6. This result is based on the analyzed users. To get the global results of the analyzed users we sum all of the positive political tendencies founded in each analyzed user. We only use 104 users out of 120, gives that we deleted users with news profiles, joking profiles, etc... which are not representative users. The other data source for contrasting the results are news related with the days that we extract the political Tweets. We select these news for checking the political occurrences that happen related to the tendency that shows the users in their Tweets.

The Figure 5.6 shows the political tendency obtained by the users analyzed TEN.

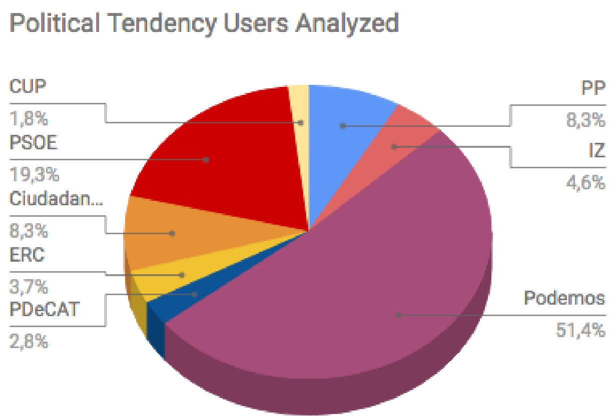

Figure 5.6: Political Tendency Users Analyzed (TEN)

5.5.1 TAS Strategy Results

In this subsection, we compare the political tendency obtained with the TAS Strategy, in relation to the results of the analyzed users (TEN).

The political tendency showed in Figure 5.7 shows the results of TAS strategy over 104 users analyzed with the assert variance between 26,83% and 72,73%, this variance depends of the number of political tweets of these users. We can appreciate that the majority of analyzed users belong to PSOE with a 51,8%, followed by Podemos with a 5,5% and with minor representation have Ciudadanos by 1,8%.

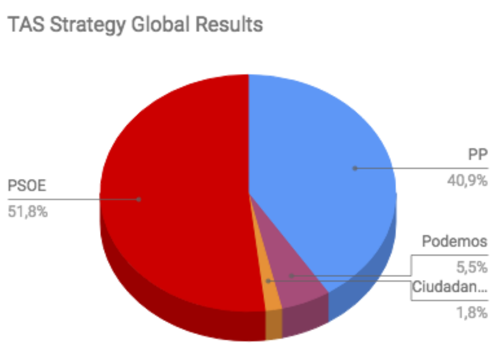

Figure 5.7: Political Tendency TAS Result

In Figure 5.8 shows the comparison between the TAS strategy results with the analyzed users results TEN. Likewise, Table 5.3 shows the mean absolute error (MAE). The MAE is a measure of forecast accuracy and has been widely used to compare the accuracy of political information markets relative to election polls (see Huber and Hauser 2005). The MAE is measured as the average of the absolute difference between the prediction and the actual result.

We can appreciate such different results, thats happen because the TAS strategy is based on analyzing the lasts users tweets and reflect the last opinion of the users and its can be conditioned by recently political occurrences. Likewise, it depends of the political moment, the users talks about some political occurrence and its reflected on the last tweets. Also the cause of the error showed in these results can be caused for the highest neutral sentiment of the analyzed tweets.

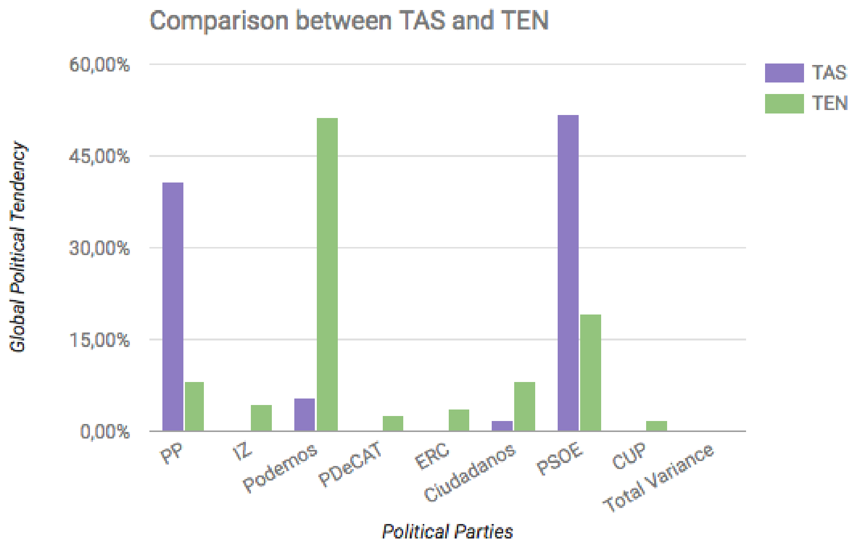

Figure 5.8: Comparison between TAS and TEN

	TAS	TEN	Prediction Error
PP	40,91%	8,26%	32,65%
IZ	0,00%	4,59%	4,59%
Podemos	5,45%	51,38%	45,92%
PDeCAT	0,00%	2,75%	2,75%
ERC	0,00%	3,67%	3,67%
Ciudadanos	1,82%	8,26%	6,44%
PSOE	51,82%	19,27%	32,55%
CUP	0,00%	1,83%	1,83%
MAE			5,51%

Table 5.3: TEN TAS Political Tendency Comparison Results

The TEN analysis strategy is similar to the TAS strategy give that TAS is the automation of the manual process carry out in the TEN. We identify that the TAS strategy is related to the political occurrences, so it is very sensitive to the occurrence of the daily news that we extract from the users Tweets. The tweet are extracted in the dates compressed between 22 of July of 2017 and 24 of July of 2017, between these days, we have some political occurrences related to PSOE [46] and PP [47]. How we can see in TAS strategy the major political tendency of the users is based on PP and PSOE, but due to the sentiment analysis give us a lot of Neutral values related to a tweets. So, we can not obtain a accurate tendency because we lost negative or positive to each parties.

In these comparison we can conclude that the TAS strategy does not get us a great political orientation but if we could improve the sentiment analysis of the tweets, we could show in real time the tendency of the users and insert information to the CAS Strategy.

5.5.2 CAS Strategy Results

We also use the set of analyzed users (TEN) to verify the CAS strategy results. Note that in this case, the TEN analysis is completely different to the CAS strategy given that the CAS strategy is based on the users friendship for getting the political tendency, while the TEN is based on analyzing the user tweets, retweets and profile contents.

The political tendency showed in Figure 5.9 shows the results about the 104 users of the TEN sample with the assert of 82,5%. We can appreciate that the majority of analyzed users are of Podemos with a 52,9%, followed by PP with a 15,4% and PSOE by 15,4% and with minor representation have Ciudadanos with a 12,5% and ERC with a 3,8%.

Figure 5.9: Political Tendency CAS Result

Likewise keeping in mind the tendency of the analyzed users (TEN) showed in Figure 5.10, we can observe a great match with the political parties, showed in Table 5.11. We use the mean absolute error (MAE).

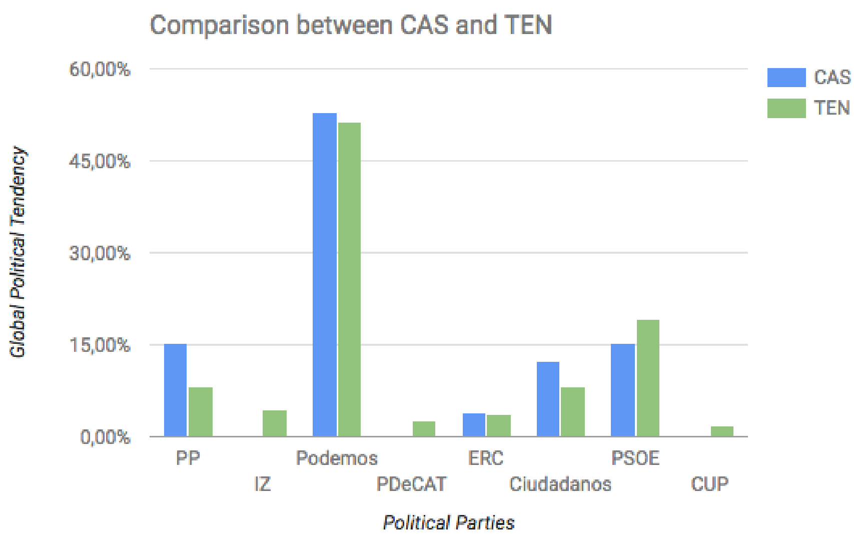

Figure 5.10: Comparison between CAS and TEN

	CAS	TEN	Prediction Error
PP	15,38%	8,26%	7,13%
IZ	0,00%	4,59%	4,59%
Podemos	52,88%	51,38%	1,51%
PDeCAT	0,00%	2,75%	2,75%
ERC	3,85%	3,67%	0,18%
Ciudadanos	12,50%	8,26%	4,24%
PSOE	15,38%	19,27%	3,88%
CUP	0,00%	1,83%	1,83%
MAE			3,32%

Figure 5.11: CAS TEN Political Tendency Comparison Results

With these results, we can conclude that the CAS strategy gives us a great way to predict the political orientation of the users, but we only can obtain a static orientation. The reason is because the relationships of the users is not a value that change frequently as opposed to the users Tweets. Users launch different opinion with her Tweets, but the relationships between the users not change frequently.

5.6 Discussion

According to the results shown in this chapter, we conclude that the best strategy to obtain the political tendency is the clustering strategy, gives that we have achieved an 82,5% of assert in front of the 67% obtained in the TAS in the best cases. Otherwise, The TAS strategy gives us the political tendency in real time in comparison with the CAS strategy that gives us the political orientation of the users. This is because in the CAS strategy we obtain the political orientation with the relationships and it does not usually change with the political occurrences. On the other hand, that with the TAS strategy, obtains the tendency in real time, because it is related with the political occurrences.

Chapter 6

Conclusions and Future work

6.1 Conclusions

This work is focused on developing a BigData platform to analyze the content of social networks with the aim to obtain the political tendency of the users. In this scope, we have worked on the specific social network, named Twitter. In this social network the content to analyze are: the tweets and the users friendship relations.

We have analyzed the Tweets content and friendship between the users. Moreover, we have researched papers related to this political topic. According to the results obtained in this study, we have defined two different strategies.

The first is named Tweet Analysis Strategy (TAS). It's focused on analyzing the users Tweets content by classifying the Tweets related to a political party and assigning a sentiment to each tweet. TAS strategy gives us the political tendency in real time because the data analysis is done with the recent launched user tweets. The result of the TAS strategy is the political tendency for each user, based on the analysis of the political tweets.

The second is named Clustering Analysis Strategy (CAS). It's focused on analyzing the relationship within the users and the relationships within the politician friendships. CAS strategy gives us the political orientation because its strategy analyzes the relationship of the users and it doesn't change frequently. In addition, it provides the interested content type of the users. Finally, it gives the community and the content type that the user is interested. The result of the CAS strategy is the political orientation for each user, based on the analysis of the users friendships.

These strategies are developed inside a BigData platform, using Storm, Hadoop and Apache Spark which are deployed inside a cloud platform named Stormy. Additionally, we have developed a simple web application to manage the system status and to visualize the results.

We check the results of each strategy with the set of analyzed users (TEN). The results of the two strategies are very grateful: in TAS strategy we have an assert of 72,73% and in CAS strategy we have an assert of 82,5%. With these results we can conclude that these strategies are useful to predict the political tendency on Twitter. Even so, we need some upgrades to give more accurated prediction. Furthermore, we need to expand to another social networks to get a more representative sample. In the next section 6.2 we introduce the possible upgrades to improve the system performance.

6.2 Future Work

With the aim to introduce the upgrades, we split them in five parts. The First one is to improve the accuracy and the performance of TAS strategy. The second one is to enhance the accuracy of CAS strategy. The third one is to mesh the TAS and CAS results in order to obtain rich information of the tendency and the changes of the users tendency. The fourth one is to improve the web application by providing a web site where we visualize and query the results, and also, we manage the system management. The fifth one is to improve the system when we launch in production.

6.2.1 Improve TAS strategy

Upgrade the assert of Sentiment Analysis

To upgrade the Sentiment Analysis we could explore to increase the positive and negative words of our Sentiment algorithm.

Upgrade Data Classification with the Hashtags related feedback

To upgrade the Data Classification module we should improve the *Collection Keywords* set. We saved the related Hashtags and users with the propose getting a different way to identify the Hashtags and Users of the *Collection Keywords*. For each user we saved in a database: the relationship within the Hashtag or the User, the sentiment, the counter of the occurrences between their relationships. In addition, we did the same for the Hashtag relationships with the related Hashtags or Users. The next step is modeling the relationships within the Hashtags and the Users with a graph and identifying new possible Hashtags and Users in order to add it in the *Collection Keywords*. These results can also provide information of the Hashtags related to politicians.

Increase Analyzed users set (TEN)

Increase the Analyzed users set (TEN) to get more results and improve the comparison in relation to our proposals.

6.2.2 Improve CAS strategy

Add politicians in clusterization of users

Investigate if the assert of the CAS algorithm will improve by adding the politicians in the users clusterization process. Likewise, we should increase the number of analyzed tweets in order to improve the accuracy.

6.2.3 Mesh CAS and TAS strategies

Investigate some correlations or information mesh to obtain new and relevant information because the CAS strategy gives us the users political orientation, and TAS strategy gives us the users political tendency in real time.

6.2.4 Improve the web management application

The next points are to enhance the web management application.

User Tendency Timeline

Show in a view the political tendency changes in a time line linked to the political occurrences. This information could be showed for each user, for each users community with a specific political orientation and for the global users.

Politicians Related to most related Hashtags

Show in a view a graph of the politicians and the more common related Hashtags and its sentiment.

CAS and TAS Analytics

Show in a view analytics of the CAS and TAS strategy working process like: Tweets analyzed, Users analyzed, last news users analyzed, etc...

Logs

Show in a view the log of errors and system blocks.

6.2.5 Improve System in production environment

Wake up the system in a production environment

Deploy into production the system with an analysis of a massive users. Add a management control for manage the errors and maintain the system every time it is awake.

Expand to another social networks

Add in the Data Extraction module other social networks like Facebook, Instagram or others.

Bibliography

[1] BARRY WELLMAN, ANABEL QUAN HAASE, JAMES WITTE, and KEITH HAMPTON. Does the internet increase, decrease, or supplement social capital? *American Behavioral Scientist*, 45(3):436–455, 2001.

[2] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 251–258, Washington, DC, USA, 2011. IEEE Computer Society.

[3] Bad results in the forecast of the elections. http://www.lavanguardia.com/politica/elecciones/20160627/402794314538/sondeos-elecciones-generales-26j-fallo.html, 2016.

[4] Andranik Tumasjan, To Sprenger, Pg Sandner, and Im Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[5] D. Sunshine Hillygus. The evolution of election polling in the united states. *Public Opinion Quarterly*, 75(5):962, 2011.

[6] David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks. *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM)*, (November 2003):556–559, 2003.

[7] Che-chia Chang. Predicting Political Affiliation of Posts on Facebook. (64), 2017.

[8] Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, and Jacob Spoelstra. Predicting us primary elections with twitter, 2012.

[9] Elvyna Tunggawan and Yustinus Eko Soelistio. And the winner is ...: Bayesian twitter-based prediction on 2016 u.s. presidential election. *CoRR*, abs/1611.00440, 2016.

[10] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment, 2010.

[11] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent Twitter Sentiment Classification. *Computational Linguistics*, pages 151–160, 2011.

[12] Yaowei Wang, Yanghui Rao, Xueying Zhan, Huijun Chen, Maoquan Luo, and Jian Yin. Sentiment and emotion classification over noisy labels. *Knowledge-Based Systems*, 111:207–216, 2016.

[13] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093 – 1113, 2014.

[14] Tianxi Li and Yu Wu. Twitter hash tag prediction algorithm. 2010.

[15] Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.

[16] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[17] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[18] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[19] Hao Zhang, Bogdan Marius Tudor, Gang Chen, and Beng Chin Ooi. Efficient in-memory data management: An analysis. *Proc. VLDB Endow.*, 7(10):833–836, June 2014.

[20] Mariam Kiran, Peter Murphy, Inder Monga, Jon Dugan, and Sartaj Singh Baveja. Lambda architecture for cost-effective batch and speed big data processing. In *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data)*, BIG DATA '15, pages 2785–2792, Washington, DC, USA, 2015. IEEE Computer Society.

[21] Mllib scalable machine learning library of spark. http://spark.apache.org/mllib/, 2017.

[22] Ankur Dave, Alekh Jindal, Li Erran Li, Reynold Xin, Joseph Gonzalez, and Matei Zaharia. Graphframes: An integrated api for mixing graph and relational queries. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems*, GRADES '16, pages 2:1–2:8, New York, NY, USA, 2016. ACM.

[23] Ankur Dave. Graphframes: Graph queries in spark sql. https://spark-summit.org/east-2016/events/graphframes-graph-queries-in-spark-sql/, 2016. retrieved at 2017-04-03.

[24] Quinton Anderson. *Storm Real-Time Processing Cookbook*. Packt Publishing, 2013.

[25] Aaron Hill, Joshua Rosslein. tweepy. https://github.com/tweepy/tweepy, 2017.

[26] Yusuke Yamamoto. Java library for the twitter api@ONLINE. http://www.twitter4j.org/, 2007.

[27] Twitter rest api. https://dev.twitter.com/rest/public, 2017.

[28] Anuchart Tassanaviboon and Guang Gong. Oauth and abe based authorization in semi-trusted cloud computing: Aauth. In *Proceedings of the Second International Workshop on Data Intensive Computing in the Clouds*, DataCloud-SC '11, pages 41–50, New York, NY, USA, 2011. ACM.

[29] A. A. Abd El-Aziz and A. Kannan. Json encryption. In *2014 International Conference on Computer Communication and Informatics*, pages 1–6, Jan 2014.

[30] Twitter api rest limitations. https://dev.twitter.com/rest/public/rate-limits, 2017.

[31] Twitter streaming api. https://dev.twitter.com/streaming/overview, 2017.

[32] Kristina Chodorow and Michael Dirolf. *MongoDB: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2010.

[33] P. Grover and R. Johari. Bcd: Bigdata, cloud computing and distributed computing. In *2015 Global Conference on Communication Technologies (GCCT)*, pages 772–776, April 2015.

[34] Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. KAF: a generic semantic annotation format. *Workshop on Semantic Annotation at GL*, 2009.

[35] Language-independent named entity recognition. http://www.cnts.ua.ac.be/conll2003/ner/, 2005.

[36] Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. Automatic creation of arabic named entity annotated corpus using wikipedia. In *EACL*, 2014.

[37] Example opener (ned) san juan. https://en.wikipedia.org/wiki/San_Juan, 2017.

[38] Dbpedia spotlight. https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki, 2017.

[39] Simple yet powerful turnkey solution to build clouds and manage data center virtualization. https://opennebula.org/, 2017.

[40] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007.

[41] Twitter api. https://dev.twitter.com/, 2017.

[42] Opener project. http://www.opener-project.eu/, 2017.

[43] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[44] Lookup twitter user information api. https://dev.twitter.com/rest/reference/get/users/lookup, 2017.

[45] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[46] News related to political party psoe. http://www.elperiodico.com/es/noticias/politica/noticias-hoy-22-mayo-2017-6052352, 2017.

[47] News related to political party pp. http://www.elperiodico.com/es/noticias/politica/rajoy-comparecera-congreso-comision-investigacion-partido-popular-6
2017.

Chapter 7

Appendix

The purpose of this section is; Present the users Analyzed TEN, that shows the users analyzed manually. Present the Tweets analyzed and the detail results of the own Sentiment Analysis strategy. Show the results of the users analyzed by the CAS strategy and we can see the result of the CAS strategy compared with the result of the users analyzed TAS. Present an example of the Twitter Streamin API output, thats shows 3 tweet examples in JSON format. Finally we can see the Keywords and Collection Keywords used in the TAS strategy.

7.1 Users Analyzed (TEN)

This section shows the results of the users analyzed, we can see a table with the user number and the analyzed political parties, the result for each political party can be -1 if we found a negative tendecy to the political party, 1 if we found a positive tendency and empty if we don't found any reference to the political party. For each row we have the analysis for each user.

User	PP	PSOE	C's	Podemos	PDE	CUP	ERC	IU
1	-1	-1		1				
2	-1	-1	-1	1				
3	-1	-1	-1		-1			
4	-1	-1	-1		-1			
5	-1	-1			-1			
6	-1	-1	-1	1				
7	-1							
8	-1	-1	-1	1				
9	-1	-1	-1	-1				
10	-1			1				
11	-1	-1		1				
12	-1	-1	-1		-1			
13	-1	1	-1	-1	-1	-1		
14	-1	-1		1				
15	-1	-1		1				
16	-1	1		1				
17	-1							
18	-1	1		1				
19	-1	-1		-1				
20	-1	-1	-1	1				
21	-1							
22	1	-1						
23	-1	-1						
24	-1	-1	-1	-1	1	1	1	
25	-1	-1		1				
26	-1	1		-1				
27	1	-1	-1	-1	-1	-1	-1	
28	-1	1		1				
29		-1	-1	-1				
30	-1	-1	-1	1				
31		1						
32	-1	-1		1				
33			1					
34		-1		-1				
35	-1			1				
36	-1							-1
37	-1		-1	1				1
38	-1	1	-1					
39	-1			1				
40	-1	-1	-1	1				
41	-1	1						
42	-1	-1	-1	1				
43	-1	-1	-1	1				
44	-1	-1		1				
45	-1	-1	-1					
46	-1		-1	1				
47	-1			-1				

Table 7.1: Users Analyzed TEN

48	-1	1		-1				
49	-1	1	-1	-1	-1	-1	-1	-1
50	-1	1	-1					
51	-1	-1	-1	1	-1	-1	-1	-1
52	-1	-1	1	-1	-1	-1	-1	
53	-1	-1	-1	1	-1	-1	-1	-1
54	-1	-1	-1	1				
55	-1	-1		1				
56	-1	-1	-1	1				
57	-1	-1	-1	1				
58	-1	-1	-1	1	-1	-1	-1	-1
59	-1	-1	-1	1				
60	-1	1	-1	-1	-1	-1	-1	-1
61			1					
62	-1	-1	1	-1				
63	1	-1	1	-1				
64	-1	-1	1	-1				
65	-1	-1		1				
66	-1	-1	-1	1				
67	-1	-1	-1	1	-1			
68	-1	1	-1	1				
69		1	1					
70		-1	-1	-1			-1	
71	1	-1	-1	-1				
72		1		-1	-1		-1	
73	-1		-1	-1				
74	-1	-1	-1	1			-1	
75	-1	-1	-1	1	-1			
76	1	-1		-1	-1		-1	
77	-1			1				1
78	-1	-1		1				
79	-1	-1	-1	1				
80	-1	-1		1				
81		-1						
82	-1	-1						
83	-1			1				
84	-1	-1	-1	1				
85	-1	-1						
86	1	-1	-1	-1	-1	-1	-1	-1
87	-1	1	-1	-1	-1	-1	-1	-1
88	1	-1		-1				
89	-1	-1	-1	1				
90	-1	-1	-1	-1				
91	-1	-1	-1	1				
92	1	-1	-1	-1	-1			
93	-1	-1						
94	-1		-1		1	1	1	
95	-1	-1		-1				
96	-1	-1	-1	1	1		1	
97	-1		-1				1	
98	-1	-1	1	-1				
99	-1	-1		1				
100	-1	1	-1	1				
101	1	-1	-1	-1	-1	-1	-1	-1
102				1				1
103	-1	1		1				1
104	-1	1	-1	-1	-1	-1	-1	-1
105	-1	-1		1				
106	-1	-1	-1	1				
107	-1	-1	-1	1				
108	-1	1	-1	1				
109	-1	-1	-1					
110	1			-1	-1	-1	-1	
111	-1	-1	-1	1				1
112	-1							
113	-1	-1	-1	1				
114	-1	-1	-1	1				
115	-1	-1	-1	1				
116	-1	1		1				
117	-1	1		1				
118	-1	-1	1	-1	-1	-1	-1	-1
119	-1	-1	-1	1				1
120	-1	-1	-1	1				

Table 7.2: Users Analyzed TEN

7.2 Sentiment Analysis

This section shows the results of the analysis of the Tweets. We can found a table and in the header we have; The % assert. The real sentiment, based on a manual analysis. The result that shows the result of the sentiment analysis algorithm.

	Assert	Real Sentiment	Result	Tweet
1	50%	-1	0	Defendiendo 29 enmiendas Presupuesto de Empleo.Ninguna apoyada por Ciudadanos/PSOE.Ellos son los responsables del… https://t.co/LT7jh1bkos
2	50%	-1	0	@GppaytoGranada denunciamos el batiburrillo de borrador de presupuestos que ha presentado el #PSOE en el Ayto.… https://t.co/jZv9wjFIBm
3	0%	-1	1	Los niños crecen y nos han salido fascistas. PP-PSOE-C's parecen del mismo partido, bueno, en realidad lo son. https://t.co/8SokK9xIfA
4	0%	-1	1	Algunos en #Parla se empeñan en hacer lo mismo para obtener los mismos resultados económicos. Apoyo al psoe antes, ahora apoyo al PP = Ruina
6	0%	-1	1	.@Pablo_Iglesias_ Por victorias así es importante recordar quiénes lleváis años pidiendo abandonar la UE para impri… https://t.co/FzmRp3yUFZ
7	50%	-1	0	Alguien le debería de decir al PSOE que la oposición es contra el PP no contra los "socialistas de a pie"… https://t.co/eT4uEwphyA
8	100%	-1	-1	"En este país las eléctricas violan los DDHH con la complicidad del Gobierno del PP y anteriores gobiernos del PSOE… https://t.co/QttgPuiLef
9	50%	-1	0	¡¡¡Se acabo la crisis!!! 8.000 Euros al mes para la Princesa Leonor - Pablo Iglesias Podemos https://t.co/7868QY6weL ja no hi ha pobresa
10	50%	-1	0	Aznar renuncia a la presidencia de honor del PP para centrarse en acabar con James Bond. https://t.co/ijKy8N8lAk
11	0%	-1	1	@anhela_ Me cuesta mucho la verdad, pero antes que a Susana voto o podemos, ella estará siempre con el pp y esta vez va a ser que no
13	50%	-1	0	@AlcaudonReal1 @carolacaracola5 El Supremo solo sirve cuando hay que dar la razón a los amigos del PP y condenar a los Ciudadanos VENDIDOS
14	100%	1	1	@Viejis39 Lo han dejado , la iniciativa de PODEMOS la harán VALIDA dentro de 3 MESES , cuando se haya IDO el FRÍO , SENSIBILIDAD PPxxE !!
15	0%	-1	1	PSOE se pone de acuerdo con PP y C'S para blindar a los torturadores del franquismo. Suena fuerte pero es verdad. https://t.co/b4oAsWR5zP
16	0%	-1	1	El PSOE votando contra el PSOE. ¿Cómo pueden dar amparo a los torturadores del franquismo? https://t.co/jjYwqVk56m
17	100%	-1	-1	El ministro de Exteriores del #PP se permite decir barbaridades, como van arreglar algo, si ni siquiera les preocupa https://t.co/lR8bcqxeXT
18	50%	1	0	El PSOE ha pactado salario mínimo y alivio de pobreza energética. Cs ampliación permiso paternidad y complemento salarial. ¿Qué hace Ps?
20	100%	1	1	En Antena-3 informaba así del ataque a la sede PP por un desequilibrado en 2.014 Pues bien, TVE lo ha puesto al fr… https://t.co/aanAgA5Atr
21	100%	1	1	@17jordani @Carpe6763Real @CiudadanosCs esto es una cadena sin fin suma y sigue
22	50%	-1	0	Así es como "trabajan" l@s diputad@s del @PPopular Háganlo ustedes en el trabajo y me cuentan..... #ClausulaSuelo https://t.co/pIeKGMl3yh
23	50%	1	0	@ahorapodemos @Pablo_Iglesias_ y luego llegarán las checas? https://t.co/XK3igK6lfN
24	50%	-1	0	@Pablo_Iglesias_ Menudo combo ETA-VENEZUELA-CUP-PODEMOS!!!!!!!! jajajjajajajaajaja
25	50%	-1	0	Que Aznar deje de ser Presidente de Honor del PP me la suda, yo lo que quiero saber cuando va a ser investigado, ta… https://t.co/qgrvlPljUt
27	50%	1	0	"Queremos garantias de que este invierno no se le va a cortar la luz y la calefacción a nadie" @Pablo_Iglesias_ #SinLuz21D
28	100%	-1	-1	PP, PSOE y Cs pactan a oscuras sobre pobreza energética. Seguimos esperando que aclaren si apoyan nuestras 6 reivin… https://t.co/mPadgNCosY
29	100%	1	1	Podemos y los "comunes" se suman a JxS y la CUP en la censura al Constitucional https://t.co/nnmfYaqCjR vía @indpcom
30	50%	1	0	Primeros movimientos del mercado de fichajes invernal: el PP pierde a su extremo derecho. https://t.co/ndZV5OJ4pj
31	100%	-1	-1	Los bancos condenados a devolver 4000m Rescatar autopistas nos cuesta 5000m Rajoy ha sacado 9000 d la hucha d las pensiones #ClausulaSuelo
32	100%	0	0	@antoine0810 ta pp tte rose la
34	50%	-1	0	@CiudadanosCs EL CASO ES ROBAR, EL COMO NO IMPORTA #STOPprivilegios #VenezuelaEP Granados #CámbiameLotería… https://t.co/TSjDXOCsxK
35	0%	-1	1	Que mejor fichaje para el PP que Álvaro Zancajo para seguir manipulando en 24 horas. Ya era un crack en Antena 3 https://t.co/jqr5Fb0psn
36	50%	1	0	Aznar deja la presidencia de honor del PP, suena Felipe González.
37	50%	-1	0	¿Cuales serán los intereses ocultos de @JM_Kichi para firmar ese contrato? Los gaditanos estarán lamentando sus vot… https://t.co/epYRBAfm0a
38	100%	-1	-1	Granados niega ante el juez una contabilidad B en el PP de Madrid y rechaza que disponga de.. https://t.co/xb03CSa5r1
39	50%	1	0	Zapatero: Hay que volver a mejorar las pensiones mínimas y recuperar la financiación de la dependencia olvidada por el PP #AlcaldesPSOE
41	50%	-1	0	Estos correveidiles han perdido la poca vergüenza que les quedaba. Hablan y escriben siguiendo la voz de su "amo". https://t.co/89gYFt1WcY
42	50%	-1	0	Esto es con lo que no están contando los golpistas, que el cabreo en vez de ir a menos va a más y les puede hacer d… https://t.co/pgaCr12jyV
43	50%	1	0	Cuando un gobierno quiere trabajar para la gente lo primero es reunirse con las organizaciones de la soc civil… https://t.co/mY9tWBq66E
44	0%	-1	1	@CRAZYLOVEHEARTS @marianorajoy yo flipo con este señor. Puede ser más paleto? Pida un cerebro por Navidad, Marco Rajao Burro :)
45	50%	-1	0	@LekaconK está cada vez más a la Derecha se confunden con el PP HASTA DESAPARECER !!!
46	50%	-1	0	A naftalina..? Mira quien fue a hablar, el comunista forgiano, decimonónico con libro y gafas de culo de vaso.... https://t.co/KQ3YWuaCXd
48	100%	1	1	El Gobierno de Rajoy aprobó 52 medidas para imponer el castellano mientras estaba en funciones… https://t.co/kIc4SqD94R
49	0%	-1	1	En Podemos todo son sonrisas, besos y diversión, hasta que alguien te pregunta: ¿Has pagado la cuota de este mes?
50	0%	-1	1	Cuando descubras la obra de su hermano Jorge VAS A FLIPAR. @marianorajoy https://t.co/tWp7OghuAS
51	50%	1	0	RT @ahorapodemos "Queremos garantias de que este invierno no se le va a cortar la luz y la calefacción a nadie" @Pablo_Iglesias_ #SinLuz21D
52	50%	-1	0	El Tribunal Europeo, el Tribunal Europeo os habéis enterado. Listillos, ni el PP ni el PSOE. El Tribunal Europeo. https://t.co/1FzLvdExh0
53	50%	1	0	RT @ahorapodemos "Queremos garantias de que este invierno no se le va a cortar la luz y la calefacción a nadie" @Pablo_Iglesias_ #SinLuz21D
55	100%	1	1	Flor vigna es todo lo que esta bien. Tan incondicional con pp desde el dia uno, tan respetuosa y buena onda. Me cerro la boca posta
56	100%	1	1	Nuestro homenaje a un genio: Paco de Lucía https://t.co/2mQO8PKZPk
57	50%	1	0	"Esta movilización forma parte de lo que nosotros llamamos oposición social y popular " @Pablo_Iglesias_ #SinLuz21D https://t.co/5FWqe2oJkJ
58	100%	-1	-1	EXCLUSIVA DE VP - #RitaGate Los proveedores del PP tuvieron que borrar de las facturas el rastro de Barberá… https://t.co/QrQKd6akSA
59	50%	1	0	Aznar renuncia a la presidencia de honor del PP para centrarse en acabar con James Bond. https://t.co/ijKy8N8lAk
60	50%	1	0	A la falta de Empleo el ministro de Exteriores le llama "Inquietud" #YoMeTiroAlMonte @PPopular https://t.co/pz2VUciGIE
62	100%	1	1	"Defendemos que no se le pueda cortar la luz o la calefacción a nadie por el hecho de ser pobre" @Pablo_Iglesias_… https://t.co/4AfPFTEGNH
63	100%	1	1	.@Albert_Rivera "No hace falta tardar 3 años en suprimir los aforamientos, se puede realizar con una reforma exprés… https://t.co/J6Pi617uAN
64	100%	1	1	"Para nosotros es un honor estar aquí acompañando a la gente" @Pablo_Iglesias_ #SinLuz21D https://t.co/8PW8lvIVrV
65	50%	1	0	Rajoy eleva a Borges a trending topic al cambiarle el nombre via @El_Plural https://t.co/T50K28q0O3 https://t.co/G2vbXfO0Hz
66	100%	-1	-1	.@CiudadanosCs acusa al Gobierno de "mentir" sobre el pacto contra la pobreza energética https://t.co/trKEjIxycB https://t.co/wuYQ4tLaYa
67	0%	-1	1	VÍDEO 15.000 euros por un vídeo electoral de Barberá, la campaña por encima de sus posibilidades del PP valenciano… https://t.co/J6UzBtmCD5
69	50%	1	0	#España Maillo: "En el @PPopular mejoramos el sistema para dar más participación… @mdcospedal @marianorajoy https://t.co/0IS13g3Meb
70	0%	-1	1	Papelón el que ha hecho hoy el PSOE no recibiendo a las víctimas del franquismo que se suma a su postura ayer respecto a la Ley de Amnistía.
71	100%	1	1	El Gobierno de Rajoy aprobó 52 medidas para imponer el castellano mientras estaba en funciones… https://t.co/kIc4SqD94R
72	0%	0	0	Marcha federal https://t.co/cO4qkKCelr
73	50%	-1	0	ESTAS RATAS SON MAS DE CELIBRAR LA NAVIDAD ARROYADO A TODALA GENTE CON UN CAMION EN LA PLAZA MAYOR https://t.co/BryCpovqxJ
74	50%	-1	0	¿Por qué Rajoy, Aguirre, Rivera... cuando hacen referencia a la literatura terminan metiendo la gamba? José Luis Borges
76	100%	-1	-1	@zoidoJI No acepta derogar #LeyMordaza ,solo reformar algún aspecto. @antoniotrevin :considere #PP en Congreso la P… https://t.co/95EyW9KeMX
77	100%	1	1	Fin de los cortes de luz a los hogares con menos recursos, objetivo acabar con la pobreza energética #QueNoTeCorten https://t.co/khM8yuvUZA
78	0%	-1	1	El ridiculo espantoso al que llegan algunos miembros del PSOE para intentar ocultar su afinidad con el PP https://t.co/kLkjdlsrj9
79	50%	-1	0	Donde esta su compasión ?? Donde esta su humanidad ? @marianorajoy @MAECgob #FREEYUYEE923 https://t.co/8Xqf2nuMIt
80	100%	1	1	@Lluis_Casanovas @MonDiari y el PP siempre ha amado a los catalanes...
81	50%	-1	0	Los podemitas de Getafe piden celebrar bautizos civiles y el PP dice que es una "chorrada" https://t.co/3avedWL9N https://t.co/O2nRNsWGEL
83	50%	-1	0	Rivera es el típico que en las mudanzas no carga un mueble pero da indicaciones y cuando acabas dice ANDA K RÁPIDO… https://t.co/nfIzmkfI6G
84	100%	-1	-1	PSC y ERC evitan rebajar más el IBI en l'Hospitalet. Rebajarlo 0,5% después de subirlo casi un 30% es un engaño …https//t.co/l5SQqSu7Zs
85	100%	-1	-1	Así se las gasta el PP: "Sanidad cesa a una enfermera de #Parla por denunciar la falta de vacunas por Twitter"…. https://t.co/yC1mFDzJ6z
86	50%	1	0	Rivera es el típico que en las mudanzas no carga un mueble pero da indicaciones y cuando acabas dice ANDA K RÁPIDO…https://t.co/nfIzmkfI6G
87	0%	-1	1	Papelón el que ha hecho hoy el PSOE no recibiendo a las víctimas del franquismo que se suma a su postura ayer respecto a la Ley de Amnistía.
88	100%	-1	-1	@a_lo_gonzo Estaban El @PSOE estaba ocupado haciéndose una foto con su socio el @PPopular. Que penita dan
90	50%	-1	1	Papelón el que ha hecho hoy el PSOE no recibiendo a las víctimas del franquismo que se suma a su postura ayer respecto a la Ley de Amnistía.
91	100%	1	1	Frena también @AdaColau la expulsión de nuestros barrios en @d_bosc son tus palabras y solo tú puedes hacerlo cita… https://t.co/yv901myDEA
92	50%	1	0	Sí no lo veo no lo creo. El PP decía que la pobreza energética no existía. Y asumen la agenda de Pablo Iglesias. https://t.co/6qWKXbmdAH
93	100%	1	1	@ahorapodemos pues llegais tarde pppsoe mas cuñao os dejan fuera eso es lo que se vera en titulares aunque no sirva para este invierno.
94	50%	-1	0	El PSOE evita recibir a víctimas del franquismo tras preguntar al Gobierno por la Ley de Memoria https://t.co/sMwysExI4x por @ATorrus
95	50%	-1	0	El PP no te merece, tía. https://t.co/wDnUoIef5k
97	0%	1	-1	@gabrielrufian Seguimos teniendo dos Españas, la de los de siempre y la de los que miran al futuro.
98	0%	-1	1	Papelón el que ha hecho hoy el PSOE no recibiendo a las víctimas del franquismo que se suma a su postura ayer respecto a la Ley de Amnistía.
99	50%	1	0	Siempre viene bien #Recordar. Nunca está de mas #Refrescar la memoria. Estos son los datos, suyas "las opiniones"... https://t.co/d2jyafpUb2
100	100%	1	1	Hemos llegado a un acuerdo con el PSOE, al que también se ha sumado C's, para .. https://t.co/agkbiILUKw by #PPopular via @0nvey
101	0%	-1	1	Que mejor fichaje para el PP que Álvaro Zancajo para seguir manipulando en 24 horas. Ya era un crack en Antena 3 https://t.co/jqr5Fb0psn

Table 7.3: Sentiment Analysis Tweets Analyzed

102	100%	-1	-1	Lo malo es que suena como si fuesen a dejar fuera a todos los que no pueden pagar la luz! https://t.co/j6yIFJ4IbK
104	100%	-1	-1	Revés del Parlamento de Flandes a Rajoy por el juicio a Carme Forcadell @forcadellcarme @lavanguardia https://t.co/tMkl3gHHgN
105	50%	1	0	Una entrevista diferente con algunas preguntas clave de @magdabandera (fascismo, establishment, izquierda, capital) https://t.co/oeFGYw8sRh
106	100%	1	1	Unidos, podemos enfrentar los retos que vienen. Nuestra prioridad será proteger e impulsar a los mexicanos donde qu…https://t.co/Yr1gJpQhRd
107	50%	-1	0	@linternacope El PSOE sigue girando muy a la izquierda. Y miedo me da ver a Jose Luiz . Zp por ahí otra vez
108	50%	-1	0	@MatiasAlonso_2 @Montseblanca @PSOE lo mismo q tos ls politicos independ de s "ideologia" vivieron y viven d rent en esta "falsa" democraci
109	50%	-1	0	#DefensaDeLasPensiones "Europa da asco y España apesta" (36FOTOS) (Hoy #20D Manifestaciones en toda España)… https://t.co/AYHZ8F8dRn
111	0%	-1	1	@jvmnt @Pablo_Iglesias_ bufffff han votado tantosssssss que igual en cinco días no les da para(tirar los votos en contra de Pablo) contar!
112	50%	-1	0	#STOPprivilegios PARA LOS LADRONES DE @CiudadanosCs QUE YA IMITAN AL PP CON UNA MAFIA NUEVA #VenezuelaEP… https://t.co/zLZPY4SEgF
113	50%	1	0	A la "diosa" se le rebelan sus "fieles" #MilitantesEnPie https://t.co/lMlbCXDpTR
114	100%	1	1	"Defendemos que no se le pueda cortar la luz o la calefacción a nadie por el hecho de ser pobre" @Pablo_Iglesias_… https://t.co/hSTt6WsQhO
115	0%	-1	1	Descubren en un bar 140 nuevos insultos a Podemos https://t.co/T21ooKIXeG
116	50%	1	0	ÚLTIMA HORA: Aznar deja la presidencia de honor del PP para dedicarse en cuerpo y alma a perseguir infieles. https://t.co/JrMe9ssVex
118	100%	1	1	"Defendemos que no se le pueda cortar la luz o la calefacción a nadie por el hecho de ser pobre" @Pablo_Iglesias_… https://t.co/HLk5tLzMDI
119	0%	-1	1	@PSOE esta tradición os salió bien en la Transición porque no teniamos Informacion ,,ahora ya no cuela https://t.co/e0Y0sNNPvI
120	100%	1	1	"Para nosotros es un honor estar aquí acompañando a la gente" @Pablo_Iglesias_ #SinLuz21D https://t.co/8PW8lvIVrV
121	50%	1	0	Ya estamos preparados en Cibeles,desde Getafe...pq no podemos permitir apagones en ninguna casa #NoMasCortesDeLuz https://t.co/FGXvvSNFVG
122	0%	-1	1	@PPopular a ver si hacemos un poco de casito a la ONU guapis https://t.co/1fDyTQZUUA
123	0%	-1	1	PSOE, el socio colaborador del PP. Ayuda con tu retuit para promocionar a Felipe González como nuevo líder del PP… https://t.co/QIVtPIBUAC
125	0%	-1	1	Papelón el que ha hecho hoy el PSOE no recibiendo a las víctimas del franquismo que se suma a su postura ayer respecto a la Ley de Amnistía.
126	100%	1	1	@Conchi_Palencia @Laspedreras Es curioso que, mientras tanto, el @PSOE se felicite x hacer cumplir algún derecho en España.
127	50%	-1	0	Aznar sacó al PP del franquismo para colocarlo en la extrema derecha. Los jóvenes del PP callan. Huele a naftalina. https://t.co/36daQykoPX
128	50%	1	0	Gobierno y PSOE cierran un acuerdo para prohibir por ley los cortes de luz https://t.co/xuGVJcF4Uf
129	100%	-1	-1	No podemos permitir que nadie se quede sin luz por no poder pagar las facturas. No se trata de caridad, se trata de… https://t.co/yEykxR82mB
130	50%	1	0	Los sanchistas reclaman a @sanchezcastejon un “gesto” para seguir adelante frente a la Gestora. Por @carmentorrres https://t.co/b8ojuuFwMd
132	100%	1	1	#TomásGómez ➡《El origen de Ciudadanos y de Podemos es idéntico...》https://t.co/IM4E5OkxU ⤵https://t.co/YI1XA4a9bf
133	100%	-1	-1	Un exalcalde del PP confirma ante el juez que el PP le pidió beneficiar a empresas donantes https://t.co/NA7BpK8Z0Q vía @El_Plural
134	0%	-1	1	@PPopular El ministro de Exteriores abochorna al Congreso: los españoles emigran “por amplitud de miras” https://t.co/uEwvQhITDU
135	50%	-1	0	@Bolsacarlosmari @CiudadanaMartaR yo confiaba en @CiudadanosCs @GirautaOficial pero parece que solo interesan ciertos colectivos.
136	50%	-1	0	El PSOE evita recibir a víctimas del franquismo tras preguntar al Gobierno por la Ley de Memoria https://t.co/sMwysExl4x por @ATorrus
137	100%	1	1	Esperemos q algún día @gabrielrufian y @aKollontai se reúnan también con las familias de los asesinados por #ETA, s… https://t.co/3D2Tlbza33
139	50%	-1	0	El PSOE evita recibir a víctimas del franquismo tras preguntar al Gobierno por la Ley de Memoria | Diario Público https://t.co/9DJBrCq1mn
140	0%	-1	1	@chulopirulo @roman_ones @CondeRoman_ones ¡Anda!No.. las cajas no llevan votos para el PP https://t.co/GfwM2v38j https://t.co/9ZPFwTIG50
141	100%	-1	-1	No podemos permitir que nadie se quede sin luz por no poder pagar las facturas. No se trata de caridad, se trata de… https://t.co/yEykxR82mB
142	50%	1	0	¡Pero ahora es cuando... ! Ya mismo tenemos una batalla que no podemos perder BRINDEMOS Y TRABAJEMOS CON ENTUSIASMO… https://t.co/NpdcVo3cAZ
143	100%	-1	-1	@CiudadanosCs esa es la verdadera democracia en españa todavia estamos muy lejos de ser una verdadera democracia
144	100%	-1	-1	La RGI se puede abordar y cambiar desde varios ángulos. El del PP es criminalizar a sus perceptores e inmigrantes. https://t.co/rwDhWVFtJM
146	100%	1	1	Todos los consellers han recibido en el Parlament una citación del TSJC será por aquello de dialogar. Y es que Rajo… https://t.co/ctIdY7vIEq
147	100%	1	1	La Ministra de Sanidad de España Dolors Montserrat pregunta a Podemos si sus propuestas para erradicar pobreza infantil son las de Venezuela
148	100%	1	1	Que mejor fichaje para el PP que Álvaro Zancajo para seguir manipulando en 24 horas. Ya era un crack en Antena 3 https://t.co/jqr5Fb0psn
149	50%	1	0	La última decisión sobre @bdncapac pone de relieve el verdadero talante del gobierno de la CUP - Podemos en #Badalona
150	0%	-1	1	La factura del gas se ha incrementado un 67% mientras crecía el porcentaje de gente que no podía pagarla sin impedi… https://t.co/K5kSR6fSgd
151	100%	0	0	@VzqzPilar @CiudadanosCs Seguro que la Sentencia ha sido Gracias a Ciudadanos. Igual que la salida diaria del Sol.
153	100%	1	1	@gonzamartell en qué te podemos ayudar?
154	0%	1	-1	@isamaran1 @InformadorVeraz @jgolpe España despertó ,dimos fuerza a @susanadiaz para liderar el @PSOE Con que pase de Z-PARO me conformo.
155	50%	-1	0	@laquiti Jaja! Sabes? Es q no me gusta q m den el mítin, y menos cantando. Mira, podemos arreglarlo...me pongo tapones y tú me vas "matando"
156	0%	-1	1	El maltratador dice que ama a su mujer, el torero que ama al toro y el PP que ama España - Revista Muy... https://t.co/rpnFT802xH
157	50%	-1	0	@ahorapodemos @MayoralRafa en venezuela la viola el gobierno que es mas grave !! Volveis a manipular!! Vividores de lo publico!!
158	0%	-1	1	Ser presidente de honor del PP es como ser presidente de la asociación anti drogas del club de fans del Chapo Guzmán.
160	50%	1	0	El diputado de ERC Gabriel Rufián protagoniza la #PortadaImposible de la semana; por @PenagosC y @vizcarrafanclub. https://t.co/EZdjfp7COU
161	0%	-1	1	El PSOE acatará sumiso y manso. Nos contará con vuestra ayuda, que es por nosotros https://t.co/65MH8Z9NRO
162	100%	1	1	Aplaudimos el acuerdo para mantener las evaluaciones en #Educación, clave para la mejora de la calidad del sistema https://t.co/dfjqclTegw
163	50%	1	0	@gerardotc ni tú ni yo claro: pero es bonito pensar que casi. https://t.co/QQNGi8IHmw @lamarea_com @TheSimpsons
164	50%	1	0	"Esta movilización forma parte de lo que nosotros llamamos oposición social y popular " @Pablo_Iglesias_ #SinLuz21D https://t.co/5FWqe2oJkJ
165	50%	-1	0	Aznar rompe con Rajoy. La carta del expresidente José María Aznar al sucesor que él mismo puso, Mariano Rajoy, es... https://t.co/sYjitpRpyU
167	50%	1	0	Os acordais cuando decian que con Manuela y Podemos se acabaria el mundo y huiria la inversion extranjera de Madrid… https://t.co/br4vjiDMHZ
168	0%	-1	1	.@Ruben_Amon @el_pais La resiliencia de Rajoy. Como pasó en Cataluña y en otras muchas ocasiones. https://t.co/hkaJWxa1tV
169	100%	-1	-1	Hacer justicia y recibir a las víctimas debe contaminar. O sea que cuanto más lejos, mejor https://t.co/YjfeumQVrO
170	0%	-1	1	Nadal destaca que Podemos no contestó a su propuesta para prohibir los cortes de luz a los hogares vulnerables. https://t.co/hghGOk36C3
171	50%	-1	0	Aznar renuncia a la presidencia de honor del PP para centrarse en acabar con James Bond. https://t.co/ijKy8N8lAk
172	50%	-1	0	Ante la falta de plazas para personas mayores es necesario un nuevo Centro de Atención Diurna #VitoriaGasteiz… https://t.co/Gt8R8k0ik
174	50%	-1	0	Sobre dirección interina PSOE: "No siempre el tiempo lo cura todo, a veces lo pudre todo" Cándido Méndez, en @20m https://t.co/W4Em6Y4Cde
175	50%	-1	0	Iglesias , Monedero , Errejón y posteriornente Garzón. Seamos Objetivos y digamos la verdad cc : @Defensagob @NATO… https://t.co/dqnS244KZn
176	100%	1	1	.@anhela_ Ese capital es intransferible, por eso y por decirnos siempre la verdad CC/@sanchezcastejon te esperamos… https://t.co/JJU3rWROSb
177	100%	1	1	Su aprobación implicaría un paso más allá de la tregua invernal al garantizar el suministro mientras se esté en sit… https://t.co/JqSP3afZXh
178	0%	-1	1	Aznar, una vez rechazada la presidencia de honor del PP, sólo le queda la de deshonor.
179	0%	-1	1	@GirautaOficial sobre la reforma de la ley del TC: DECISIONISMO JURÍDICO por el "elegido iluminado" ,no DEMOCRÁCIA https://t.co/2PfYt9Gw6d
181	50%	-1	0	Tuiteros militantes y simpatizantes de PP y C'S amenazan, humillan e insultan a la víctima del atentado en Berlín... https://t.co/mMpC68mk4L
182	0%	-1	1	Nadal destaca que Podemos no contestó a su propuesta para prohibir los cortes de luz a los hogares vulnerables. https://t.co/hghGOk36C3
183	100%	1	1	#VenezuelaEP lo que busca Podemos para España https://t.co/SiXEuVhLpX

Table 7.4: Sentiment Analysis Tweets Analyzed

7.3 CAS strategy users results

In this section we present the results obtained of the CAS strategy. We present the results with a table where the header is composed by the user, Algorithm result (shows the main result of CAS strategy) and positive, negative and neutral parties shows the tendency of the TEN analyzed users. For each row we have the analysis for each user.

User	Algorithm Result	Positive Party	Negative Party	Neutral Party
1	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
2	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
3	Podemos		PP,PSOE,C's,PDE	Podemos,CUP,ERC,IZ
4	Podemos		PP,PSOE,C's,PDE	Podemos,CUP,ERC,IZ
5	Podemos		PP,PSOE,PDE	C's,Podemos,CUP,ERC,IZ
6	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
7	Podemos		PP	PSOE,C's,Podemos,PDE,CUP,ERC,IZ
8	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
9	Podemos		PP,PSOE,C's,Podemos	PDE,CUP,ERC,IZ
10	PP	Podemos	PP	PSOE,C's,PDE,CUP,ERC,IZ
11	C's	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
12	Podemos		PP,PSOE,C's,PDE	Podemos,CUP,ERC,IZ
13	Podemos	PSOE	PP,C's,Podemos,PDE,CUP	ERC,IZ
14	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
15	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
16	PSOE	PSOE,Podemos	PP	C's,PDE,CUP,ERC,IZ
17	Podemos		PP	PSOE,C's,Podemos,PDE,CUP,ERC,IZ
18	PSOE	PSOE,Podemos	PP	C's,PDE,CUP,ERC,IZ
19	PP		PP,PSOE,Podemos	C's,PDE,CUP,ERC,IZ
20	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
21	PSOE		PP	PSOE,C's,Podemos,PDE,CUP,ERC,IZ
22	ehbildu			PP,PSOE,C's,Podemos,PDE,CUP,ERC,IZ
23	PP	PP	PSOE	C's,Podemos,PDE,CUP,ERC,IZ
24	ERC		PP,PSOE	C's,Podemos,PDE,CUP,ERC,IZ
25	C's			PP,PSOE,C's,Podemos,PDE,CUP,ERC,IZ
26	ERC	PDE,CUP,ERC	PP,PSOE,C's,Podemos	IZ
27	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
28	C's	PSOE	PP,Podemos	C's,PDE,CUP,ERC,IZ
29	PP	PP	PSOE,C's,Podemos,PDE,CUP,ERC	IZ
30	Podemos	PSOE,Podemos	PP	C's,PDE,CUP,ERC,IZ
31	PP		PSOE,C's,Podemos	PP,PDE,CUP,ERC,IZ
32	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
33	PSOE	PSOE		PP,C's,Podemos,PDE,CUP,ERC,IZ
34	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
35	C's	C's		PP,PSOE,Podemos,PDE,CUP,ERC,IZ
36	PP		PSOE,Podemos	PP,C's,PDE,CUP,ERC,IZ
37	Podemos	Podemos	PP	PSOE,C's,PDE,CUP,ERC,IZ
38	Podemos		PP,IZ	PSOE,C's,Podemos,PDE,CUP,ERC
39	Podemos	Podemos,IZ	PP,C's	PSOE,PDE,CUP,ERC
40	PSOE	PSOE	PP,C's	Podemos,PDE,CUP,ERC,IZ
41	PP	Podemos	PP	PSOE,C's,PDE,CUP,ERC,IZ
42	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
43	PSOE	PSOE	PP	C's,Podemos,PDE,CUP,ERC,IZ
44	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
45	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
46	Podemos	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
47	Podemos		PP,PSOE,C's	Podemos,PDE,CUP,ERC,IZ
48	C's	Podemos	PP,C's	PSOE,PDE,CUP,ERC,IZ
49	PSOE		PP,Podemos	PSOE,C's,PDE,CUP,ERC,IZ
50	PSOE	PSOE	PP,Podemos	C's,PDE,CUP,ERC,IZ
51	PSOE	PSOE	PP,C's,Podemos,PDE,CUP,ERC,IZ	
52	PP	PSOE	PP,C's	Podemos,PDE,CUP,ERC,IZ
53	Podemos	Podemos	PP,PSOE,C's,PDE,CUP,ERC,IZ	
54	C's	C's	PP,PSOE,Podemos,PDE,CUP,ERC	IZ
55	Podemos	Podemos	PP,PSOE,C's,PDE,CUP,ERC,IZ	
56	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
57	PSOE	Podemos	PP,PSOE	C's,PDE,CUP,ERC,IZ
58	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
59	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
60	Podemos	Podemos	PP,PSOE,C's,PDE,CUP,ERC,IZ	
61	Podemos	Podemos	PP,PSOE,C's	PDE,CUP,ERC,IZ
62	PSOE	PSOE	PP,C's,Podemos,PDE,CUP,ERC,IZ	
63	C's	C's		PP,PSOE,Podemos,PDE,CUP,ERC,IZ
64	C's	C's	PP,PSOE,Podemos	PDE,CUP,ERC,IZ
65	PP	PP,C's	PSOE,Podemos	PDE,CUP,ERC,IZ

Table 7.5: CAS strategy users results

66	PSOE	Cs	PP,PSOE,Podemos	PDE,CUP,ERC,IZ
67	Podemos	Podemos	PP,PSOE	Cs,PDE,CUP,ERC,IZ
68	PP	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
69	Podemos	Podemos	PP,PSOE,Cs,PDE	CUP,ERC,IZ
70	Podemos	PSOE,Podemos	PP,Cs	PDE,CUP,ERC,IZ
71	Cs	PSOE,Cs		PP,Podemos,PDE,CUP,ERC,IZ
72	Cs		PSOE,Cs,Podemos,ERC	PP,PDE,CUP,IZ
73	PP	PP	PSOE,Cs,Podemos	PDE,CUP,ERC,IZ
74	Cs	PSOE	Podemos,PDE,ERC	PP,Cs,CUP,IZ
75	Cs		PP,Cs,Podemos	PSOE,PDE,CUP,ERC,IZ
76	PSOE	Podemos	PP,PSOE,Cs,ERC	PDE,CUP,IZ
77	PSOE	Podemos	PP,PSOE,Cs,PDE	CUP,ERC,IZ
78	Cs	PP	PSOE,Podemos,PDE,ERC	Cs,CUP,IZ
79	Podemos	Podemos,IZ	PP	PSOE,Cs,PDE,CUP,ERC
80	Podemos	Podemos	PP,PSOE	Cs,PDE,CUP,ERC,IZ
81	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
82	Podemos	Podemos	PP,PSOE	Cs,PDE,CUP,ERC,IZ
83	Podemos		PSOE	PP,Cs,Podemos,PDE,CUP,ERC,IZ
84	Cs		PP,PSOE	Cs,Podemos,PDE,CUP,ERC,IZ
85	Podemos	Podemos	PP	PSOE,Cs,PDE,CUP,ERC,IZ
86	Cs	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
87	Podemos		PP,PSOE	Cs,Podemos,PDE,CUP,ERC,IZ
88	PP	PP	PSOE,Cs,Podemos,PDE,CUP,ERC,IZ	
89	PSOE	PSOE	PP,Cs,Podemos,PDE,CUP,ERC,IZ	
90	PP	PP	PSOE,Podemos	Cs,PDE,CUP,ERC,IZ
91	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
92	Cs		PP,PSOE,Cs,Podemos	PDE,CUP,ERC,IZ
93	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
94	PP	PP	PSOE,Cs,Podemos,PDE	CUP,ERC,IZ
95	Podemos		PP,PSOE	Cs,Podemos,PDE,CUP,ERC,IZ
96	ERC	PDE,CUP,ERC	PP,Cs	PSOE,Podemos,IZ
97	Podemos		PP,PSOE,Podemos	Cs,PDE,CUP,ERC,IZ
98	Podemos	Podemos,PDE,ERC	PP,PSOE,Cs	CUP,IZ
99	PP	ERC	PP,Cs	PSOE,Podemos,PDE,CUP,IZ
100	Cs	Cs	PP,PSOE,Podemos	PDE,CUP,ERC,IZ
101	PSOE	Podemos	PP,PSOE	Cs,PDE,CUP,ERC,IZ
102	Podemos	PSOE,Podemos	PP,Cs	PDE,CUP,ERC,IZ
103	PP	PP	PSOE,Cs,Podemos,PDE,CUP,ERC,IZ	
104	Podemos	Podemos,IZ		PP,PSOE,Cs,PDE,CUP,ERC
105	Podemos	PSOE,Podemos,IZ	PP	Cs,PDE,CUP,ERC
106	PSOE	PSOE	PP,Cs,Podemos,PDE,CUP,ERC,IZ	
107	Podemos	Podemos	PP,PSOE	Cs,PDE,CUP,ERC,IZ
108	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
109	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
110	Podemos	PSOE,Podemos	PP,Cs	PDE,CUP,ERC,IZ
111	Podemos		PP,PSOE,Cs	Podemos,PDE,CUP,ERC,IZ
112	PP	PP	Podemos,PDE,CUP,ERC	PSOE,Cs,IZ
113	ERC	Podemos,IZ	PP,PSOE,Cs	PDE,CUP,ERC
114	ERC		PP	PSOE,Cs,Podemos,PDE,CUP,ERC,IZ
115	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
116	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
117	Podemos	Podemos	PP,PSOE,Cs	PDE,CUP,ERC,IZ
118	Podemos	PSOE,Podemos	PP	Cs,PDE,CUP,ERC,IZ
119	Podemos	PSOE,Podemos	PP	Cs,PDE,CUP,ERC,IZ
120	Podemos	Cs	PP,PSOE,Podemos,PDE,CUP,ERC,IZ	

Table 7.6: CAS strategy users results

7.4 Twitter Streaming API output example

In this section we can see an example of three tweets with its attributes.


```

/* 1 */
{
    "_id" : NumberLong(797911426950168576),
    "createdAt" : "Nov 13, 2016 10:18:03 PM",
    "text" : "RT @QuiqueVal: - Enchufan 6 asesores mas llegando al limite legal\n- Que asco Rajoy\n- Es Compromis \n- Calla facha, son los asesores ",
    "source" : "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "isTruncated" : false,
```

```
    "inReplyToStatusId" : -1,
    "inReplyToUserId" : -1,
    "isFavorited" : false,
    "isRetweeted" : false,
    "favoriteCount" : 0,
    "retweetCount" : 0,
    "isPossiblySensitive" : false,
    "lang" : "es",
    "contributorsIDs" : [],
    "retweetedStatus" : {
        "createdAt" : "Nov 12, 2016 10:05:56 AM",
        "id" : NumberLong(797364795766935552),
        "text" : "- Enchufan 6 asesores mas llegando al limite legal \n- Que asco Rajoy\n- Es Compromis \n- Calla facha, son los asesores https://t.co/IlBs5vwcUd",
        "source" : "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
        "isTruncated" : true,
        "inReplyToStatusId" : -1,
        "inReplyToUserId" : -1,
        "isFavorited" : false,
        "isRetweeted" : false,
        "favoriteCount" : 165,
        "retweetCount" : 337,
        "isPossiblySensitive" : false,
        "lang" : "es",
        "contributorsIDs" : [],
        "userMentionEntities" : [],
        "urlEntities" : [
            {
                "url" : "https://t.co/IlBs5vwcUd",
                "expandedURL" : "https://twitter.com/i/web/status/797364795766935552",
                "displayURL" : "twitter.com/i/web/status/7",
                "start" : 117,
                "end" : 140
            }
        ],
        "hashtagEntities" : [],
        "mediaEntities" : [],
        "symbolEntities" : [],
        "currentUserRetweetId" : -1,
        "user" : {
            "id" : 249756250,
            "name" : "Enrique Martinez",
            "screenName" : "QuiqueVal",
            "location" : "Valencia",
            "description" : "Ideas, trabajo, sinceridad y humildad. Licenciado en Publicidad y RRPP y Periodismo.",
            "descriptionURLEntities" : [],
            "isContributorsEnabled" : false,
```

```
            "profileImageUrl" : "http://pbs.twimg.com/profile_images/749900579292454912/IiB1arcE_normal.jpg",
            "profileImageUrlHttps" : "https://pbs.twimg.com/profile_images/749900579292454912/IiB1arcE_normal.jpg",
            "url" : "http://ladoblemoral.com/",
            "isProtected" : false,
            "followersCount" : 5636,
            "profileBackgroundColor" : "C0DEED",
            "profileTextColor" : "333333",
            "profileLinkColor" : "1DA1F2",
            "profileSidebarFillColor" : "DDEEF6",
            "profileSidebarBorderColor" : "C0DEED",
            "profileUseBackgroundImage" : true,
            "showAllInlineMedia" : false,
            "friendsCount" : 2538,
            "createdAt" : "Feb 9, 2011 7:23:17 PM",
            "favouritesCount" : 186,
            "utcOffset" : -28800,
            "timeZone" : "Pacific Time (US & Canada)",
            "profileBackgroundImageUrl" : "http://abs.twimg.com/images/themes/theme1/bg.png",
            "profileBackgroundImageUrlHttps" : "https://abs.twimg.com/images/themes/theme1/bg.png",
            "profileBannerImageUrl" : "https://pbs.twimg.com/profile_banners/249756250/1458727593",
            "profileBackgroundTiled" : false,
            "lang" : "es",
            "statusesCount" : 31561,
            "isGeoEnabled" : true,
            "isVerified" : false,
            "translator" : false,
            "listedCount" : 111,
            "isFollowRequestSent" : false
        }
    },
    "userMentionEntities" : [
        {
            "name" : "Enrique Martinez",
            "screenName" : "QuiqueVal",
            "id" : 249756250,
            "start" : 3,
            "end" : 13
        }
    ],
    "urlEntities" : [
        {
            "url" : "",
            "expandedURL" : "",
            "displayURL" : "",
            "start" : 132,
```

```
            "end" : 132
        }
    ],
    "hashtagEntities" : [],
    "mediaEntities" : [],
    "symbolEntities" : [],
    "currentUserRetweetId" : -1,
    "user_id" : 1444520108
}
/* 2 */
{
    "_id" : NumberLong(797911427491057664),
    "createdAt" : "Nov 13, 2016 10:18:03 PM",
    "text" : "RT @kharzaq: El mejor argumento contra el capitalismo
       es el cambio climatico.Naomi Klein @pardodevera @radiocable
       @Pablo_Iglesias_ @agarz",
    "source" : "<a href=\"http://twitter.com/download/android\" rel
       =\"nofollow\">Twitter for Android</a>",
    "isTruncated" : false,
    "inReplyToStatusId" : -1,
    "inReplyToUserId" : -1,
    "isFavorited" : false,
    "isRetweeted" : false,
    "favoriteCount" : 0,
    "retweetCount" : 0,
    "isPossiblySensitive" : false,
    "lang" : "es",
    "contributorsIDs" : [],
    "retweetedStatus" : {
        "createdAt" : "Nov 13, 2016 11:44:59 AM",
        "id" : NumberLong(797752109630418944),
        "text" : "El mejor argumento contra el capitalismo es el
           cambio climatico.Naomi Klein @pardodevera @radiocable
           @Pablo_Iglesias_ @agarzon",
        "source" : "<a href=\"http://twitter.com/download/android\"
           rel=\"nofollow\">Twitter for Android</a>",
        "isTruncated" : false,
        "inReplyToStatusId" : -1,
        "inReplyToUserId" : -1,
        "isFavorited" : false,
        "isRetweeted" : false,
        "favoriteCount" : 27,
        "retweetCount" : 32,
        "isPossiblySensitive" : false,
        "lang" : "es",
        "contributorsIDs" : [],
        "userMentionEntities" : [
            {
                "name" : "Ana Pardo de Vera P.",
                "screenName" : "pardodevera",
                "id" : 107153756,
```

```
                "start" : 78,
                "end" : 90
            },
            {
                "name" : "Fernando Berlin",
                "screenName" : "radiocable",
                "id" : 11713422,
                "start" : 91,
                "end" : 102
            },
            {
                "name" : "Pablo Iglesias",
                "screenName" : "Pablo_Iglesias_",
                "id" : 158342368,
                "start" : 103,
                "end" : 119
            },
            {
                "name" : "Alberto Garzon",
                "screenName" : "agarzon",
                "id" : 11904592,
                "start" : 120,
                "end" : 128
            }
        ],
        "urlEntities" : [],
        "hashtagEntities" : [],
        "mediaEntities" : [],
        "symbolEntities" : [],
        "currentUserRetweetId" : -1,
        "user" : {
            "id" : NumberLong(4210939763),
            "name" : "German Iglesias",
            "screenName" : "kharzaq",
            "location" : "Santa Cruz de Tenerife, Espana",
            "description" : "Sociologo independiente. Lector empedernido y escritor ocasional.",
            "descriptionURLEntities" : [],
            "isContributorsEnabled" : false,
            "profileImageUrl" : "http://pbs.twimg.com/profile_images/666650087230189568/BQIgZ7Wq_normal.jpg",
            "profileImageUrlHttps" : "https://pbs.twimg.com/profile_images/666650087230189568/BQIgZ7Wq_normal.jpg",
            "isProtected" : false,
            "followersCount" : 73,
            "profileBackgroundColor" : "000000",
            "profileTextColor" : "000000",
            "profileLinkColor" : "9266CC",
            "profileSidebarFillColor" : "000000",
            "profileSidebarBorderColor" : "000000",
```

```
            "profileUseBackgroundImage" : false,
            "showAllInlineMedia" : false,
            "friendsCount" : 182,
            "createdAt" : "Nov 17, 2015 5:03:03 PM",
            "favouritesCount" : 2789,
            "utcOffset" : -1,
            "profileBackgroundImageUrl" : "http://abs.twimg.com/
               images/themes/theme1/bg.png",
            "profileBackgroundImageUrlHttps" : "https://abs.twimg.
               com/images/themes/theme1/bg.png",
            "profileBannerImageUrl" : "https://pbs.twimg.com/
               profile_banners/4210939763/1447777570",
            "profileBackgroundTiled" : false,
            "lang" : "es",
            "statusesCount" : 2994,
            "isGeoEnabled" : false,
            "isVerified" : false,
            "translator" : false,
            "listedCount" : 22,
            "isFollowRequestSent" : false
        }
    },
    "userMentionEntities" : [
        {
            "name" : "German Iglesias",
            "screenName" : "kharzaq",
            "id" : NumberLong(4210939763),
            "start" : 3,
            "end" : 11
        },
        {
            "name" : "Ana Pardo de Vera P.",
            "screenName" : "pardodevera",
            "id" : 107153756,
            "start" : 91,
            "end" : 103
        },
        {
            "name" : "Fernando Berlin",
            "screenName" : "radiocable",
            "id" : 11713422,
            "start" : 104,
            "end" : 115
        },
        {
            "name" : "Pablo Iglesias",
            "screenName" : "Pablo_Iglesias_",
            "id" : 158342368,
            "start" : 116,
            "end" : 132
        }
```

```
    ],
    "urlEntities" : [],
    "hashtagEntities" : [],
    "mediaEntities" : [],
    "symbolEntities" : [],
    "currentUserRetweetId" : -1,
    "user_id" : NumberLong(793129335859314688)
}
/* 3 */
{
    "_id" : NumberLong(797911422281711616),
    "createdAt" : "Nov 13, 2016 10:18:02 PM",
    "text" : "RT @Albert_Rivera: Ya no les basta con atacar nuestra sede 6 veces. Ahora agresion. Animo companeros, gracias por ser la cara democrat ",
    "source" : "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "isTruncated" : false,
    "inReplyToStatusId" : -1,
    "inReplyToUserId" : -1,
    "isFavorited" : false,
    "isRetweeted" : false,
    "favoriteCount" : 0,
    "retweetCount" : 0,
    "isPossiblySensitive" : false,
    "lang" : "es",
    "contributorsIDs" : [],
    "retweetedStatus" : {
        "createdAt" : "Nov 13, 2016 5:00:18 PM",
        "id" : NumberLong(797831465161674752),
        "text" : "Ya no les basta con atacar nuestra sede 6 veces. Ahora agresion. Animo companeros, gracias por ser la cara democrat https://t.co/qV7QtQWP86",
        "source" : "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
        "isTruncated" : true,
        "inReplyToStatusId" : -1,
        "inReplyToUserId" : -1,
        "isFavorited" : false,
        "isRetweeted" : false,
        "favoriteCount" : 797,
        "retweetCount" : 993,
        "isPossiblySensitive" : false,
        "lang" : "es",
        "contributorsIDs" : [],
        "userMentionEntities" : [],
        "urlEntities" : [
            {
                "url" : "https://t.co/qV7QtQWP86",
                "expandedURL" : "https://twitter.com/i/web/status/797831465161674752",
```

```
                "displayURL" : "twitter.com/i/web/status/7",
                "start" : 117,
                "end" : 140
            }
        ],
        "hashtagEntities" : [],
        "mediaEntities" : [],
        "symbolEntities" : [],
        "currentUserRetweetId" : -1,
        "user" : {
            "id" : 108994652,
            "name" : "Albert Rivera",
            "screenName" : "Albert_Rivera",
            "location" : "Espana",
            "description" : "Perfil oficial de Albert Rivera. Presidente de Ciudadanos. Imposible es solo una opinion.",
            "descriptionURLEntities" : [],
            "isContributorsEnabled" : false,
            "profileImageUrl" : "http://pbs.twimg.com/profile_images/739530776862203905/8iDBTHeh_normal.jpg",
            "profileImageUrlHttps" : "https://pbs.twimg.com/profile_images/739530776862203905/8iDBTHeh_normal.jpg",
            "url" : "http://www.ciudadanos-cs.org",
            "isProtected" : false,
            "followersCount" : 705328,
            "profileBackgroundColor" : "C0DEED",
            "profileTextColor" : "333333",
            "profileLinkColor" : "FA743E",
            "profileSidebarFillColor" : "D8DAFC",
            "profileSidebarBorderColor" : "FFFFFF",
            "profileUseBackgroundImage" : true,
            "showAllInlineMedia" : false,
            "friendsCount" : 2204,
            "createdAt" : "Jan 27, 2010 5:57:22 PM",
            "favouritesCount" : 8276,
            "utcOffset" : 3600,
            "timeZone" : "Madrid",
            "profileBackgroundImageUrl" : "http://pbs.twimg.com/profile_background_images/434363038028689409/canMJjhT.jpeg",
            "profileBackgroundImageUrlHttps" : "https://pbs.twimg.com/profile_background_images/434363038028689409/canMJjhT.jpeg",
            "profileBannerImageUrl" : "https://pbs.twimg.com/profile_banners/108994652/1465732500",
            "profileBackgroundTiled" : false,
            "lang" : "es",
            "statusesCount" : 44207,
            "isGeoEnabled" : true,
```

```
            "isVerified" : true,
            "translator" : false,
            "listedCount" : 4363,
            "isFollowRequestSent" : false
        }
    },
    "userMentionEntities" : [
        {
            "name" : "Albert Rivera",
            "screenName" : "Albert_Rivera",
            "id" : 108994652,
            "start" : 3,
            "end" : 17
        }
    ],
    "urlEntities" : [
        {
            "url" : "",
            "expandedURL" : "",
            "displayURL" : "",
            "start" : 136,
            "end" : 136
        }
    ],
    "hashtagEntities" : [],
    "mediaEntities" : [],
    "symbolEntities" : [],
    "currentUserRetweetId" : -1,
    "user_id" : 171432821
}
```


7.5 Keywords

In this section we can see the keywords used for extract Political Tweets *C*.

@GirautaOficial
#AlbertRivera
@Albert_Rivera
@CiudadanosCs
#RutaCiudadana
#ConIlusion
@sdelcampocs
#Ilusión
#Ciudadanos
@InesArrimadas
#AlbertPresidente
C's
Albert_Rivera
@ConvergenciaCAT
@DemocratesCAT
@reagrupament
#possible
@20dl_cat
@joseprull
@joanbague
@peresalo68
@Ferran_Bel
@franceschoms
@ehbildu
#BilduErabakira
#EHBildu
@ehbildu_legebil
ERC
#SomRepública
@Esquerra_ERC
@GabrielRufian
@JoanTarda
@junqueras
@MartaRovira
#UNPAISCONTIGO
@ahorapodemos
#Un6Dcontigo
#6DHagamosHistoria
@Pablo_Iglesias_
@AdaColau
@VickyRosell
#LeyDeImpunidad
Podemos
Pablo Iglesias
partidopopular
partido popular
pp
#PP
#EspañaEnSerio

Figure 7.1: Example Keyword set K used in the project

7.6 Collection Keywords

In this section we can see the keywords used for classify Political Tweets C..

Collection Keywords

C_1 Partido Popular

partidopopular
pp
#EspañaEnSerio
#PP
#YoVotoPP
@marianorajoy
@AlfonsoAlonsoPP
@PPopular
@Sorayapp
@mdcospedal

C_2 PSOE

PSOE
#FemForaRajoy
#SomLaSolucio
#PSOE
#OrgulloSocialista
#VOTAPSOE
#PedroPresidente
@sanchezcastejon
@PSOE
@miqueliceta

C_3 Ciudadanos

@GirautaOficial"
#AlbertRivera"
@Albert_Rivera
@CiudadanosCs
#RutaCiudadana
#Conllusion
@sdelcampocs
#Ilusión
#Ciudadanos
@InesArrimadas
#AlbertPresidente
C's
Albert_Rivera

C_4 Podemos

#UNPAISCONTIGO
@ahorapodemos
#Un6Dcontigo
#6DHagamosHistoria
@Pablo_Iglesias
@AdaColau
@VickyRosell
#LeyDelmpunidad
Podemos
Pablo Iglesias

C_5 PDeCat

@ConvergenciaCAT
@DemocratesCAT
@reagrupament
#possible
@20dl_cat
@joseprull
@joanbague
@peresalo68
@Ferran_Bel
@franceschoms

C_6 CUP

@cupnacional
@albertbotran
@CUPBerga
@CUPSabadell
@CUPTarragona
@AnnaGaSabate
#CUP
@Nudenu

C_7 ERC

ERC
#SomRepública
@Esquerra_ERC
@GabrielRufian
@JoanTarda
@junqueras
@MartaRovira

C_8 IU

#PorUnNuevoPaís
@MailloAntonio
@iuandalucia
#IU
#IzquierdaUnida
@EUIB

Figure 7.2: Example Collection Keyword set K used in the project